DOCUMENT RESUME

ED 168 831 SE 026 897

AUTHOR Abelson, Philip H., Ed.; Hammond, Allen L., Ed.
TITLE Electronics: The Continuing Revolution. No. 5 in a Series of Special "Science" Compendia.
INSTITUTION American Association for the Advancement of Science, Washington, D.C.
REPORT NO AAAS-Pub-77-4
PUB DATE 77
NOTE 221p.; For related documents, see SE 026 896-898; Not available in hard copy due to marginal legibility of original document
AVAILABLE FROM American Association for the Advancement of Science, 1515 Massachusetts Ave., N.W., Washington, D.C. 20005 ($4.50)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Computers; Computer Science; *Economics; Educational Technology; *Electronics; Electronics Industry; Industry; Marketing; Medicine; *Research; *Technological Advancement; Technology

ABSTRACT

Presented are articles originally published in "Science" during 1975-77. This issue is divided into five parts. The first part presents four articles designed to provide an overview of the past, present, and future of electronics. The second part presents twelve articles showing the pervasiveness of electronics in banking and marketing, medicine, communications, research, education, national defense, and industry. The third part discusses some policy problems resulting from the electronic revolution. These include problems of international trade, telecommunication and government regulation, and employment. The fourth part deals with computers, computer networks, microprocessors, and what computers mean to man and society. The last part, Research Frontiers, covers such areas as new memory technologies, fiber optics for telecommunications, and the limits of solid-state electronics. (BB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Grayce A. Finger

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

No. 5 in a series of special SCIENCE Compendia

Electronics: The Continuing Revolution

ED168831

American Association for the Advancement of Science

SE 026 897

Copyright © 1977 by the
American Association for the Advancement of Science
Washington, D.C. 20005

All rights reserved

Library of Congress Catalog Number 77-77257

Printed in the United States of America

The material in this book originally appeared in *Science*, the weekly journal of the American Association for the Advancement of Science. Many of the articles were first published in a special issue devoted to electronics, dated 18 March 1977. Other articles were published in other issues over a period ranging from August 1975 to December 1976.

International Standard Book Number (CB) 0-87168-217-6; (PB) 0-87168-230-3
AAAS Publication 77-4

Orders for this publication should be sent to:

AMERICAN ASSOCIATION for the ADVANCEMENT of SCIENCE
1515 Massachusetts Avenue, N.W.
Washington, D.C. 20005

Table of Contents

INTRODUCTION TO THE CONTINUING REVOLUTION

The Electronics Revolution: *P. H. Abelson* and *A. L. Hammond* 3

Electronics: Past, Present, and Future: *J. R. Pierce* 8

Evolution of Computers and Computing: *R. M. Davis* 12

Large-Scale Integration: What Is Yet to Come?: *R. N. Noyce* 18

Intellectual and Economic Fuel for the Electronics Revolution: *J. G. Linvill* and *C. L. Hogan* 23

THE PERVASIVENESS OF ELECTRONICS

Computers in Banking and Marketing: *F. E. Balderston, J. M. Carman, A. C. Hoggatt* 31

The Impact of Integrated Electronics in Medicine: *R. L. White* and *J. D. Meindl* 35

Medical Electronics (I): Computerized X-ray Scanners: *A. L. Robinson* 41

Medical Electronics (II): Computerized Scanner Explosion: *A. L. Robinson* 44

Satellite Communications: *B. I. Edelson* and *L. Pollack* 47

Computers and Research: *W. O. Baker* et al. 56

Tutored Videotape Instruction: A New Use of Electronics Media in Education: *J. F. Gibbons, W. R. Kincheloe, K. S. Down* 61

Educational Uses of the PLATO Computer System: *S. G. Smith* and *B. A. Sherwood* 68

Impact of the Electronics Revolution on Industrial Process Control: *L. B. Evans* 76

Electronics and National Defense: A Case Study: *G. P. Dinneen* and *F. C. Frick* 81

Some Recent and Future Automotive Electronic Developments: *T. O. Jones* 86

Electronic Mail: *R. J. Potter* 90

POLICY PROBLEMS

The Convergence of Computing and Telecommunications Systems: *D. Farber* and *P. Baran* 96

The Information Economy and Public Policy: *M. R. Irwin* and *S. C. Johnson* 100

International Trade in Electronics: U.S.—Japan Competition: *J. Walsh* 105

Impact of Electronics on Employment: Productivity and Displacement Effects: *A. L. Robinson* 109

COMPUTERS AND PEOPLE

What Computers Mean for Man and Society: *H. A. Simon* 116

Trends in Computers and Computing: The Information Utility: *S. E. Madnick* 121

Distributive Computer Networking: Making It Work on a Regional Basis: *R. W. Cornew* and *P. M. Morse* . . 129

Microprocessors?—An End User's View: *R. E. Dessy* 138

Microprocessor Application: A Less Sophisticated Approach: *J. T. Arnold* 146

Humanizing Computerized Information Systems: *T. D. Sterling* 151

Software Engineering: *H. D. Mills* 155

Human Performance Considerations in Complex Systems: *H. O. Holt* and *F. L. Stevenson* 161

RESEARCH FRONTIERS

Photons in Fibers for Telecommunication: *S. E. Miller* 167

New Technologies for Signal Processing: *R. W. Brodersen* and *R. M. White* 172

New Memory Technologies: *J. A. Rajchman* 179

Microelectronics: Lithographic Technologies Progress: *A. L. Robinson* 186

Mathematics and Computer Science: Coping with Finiteness: *D. E. Knuth* 189

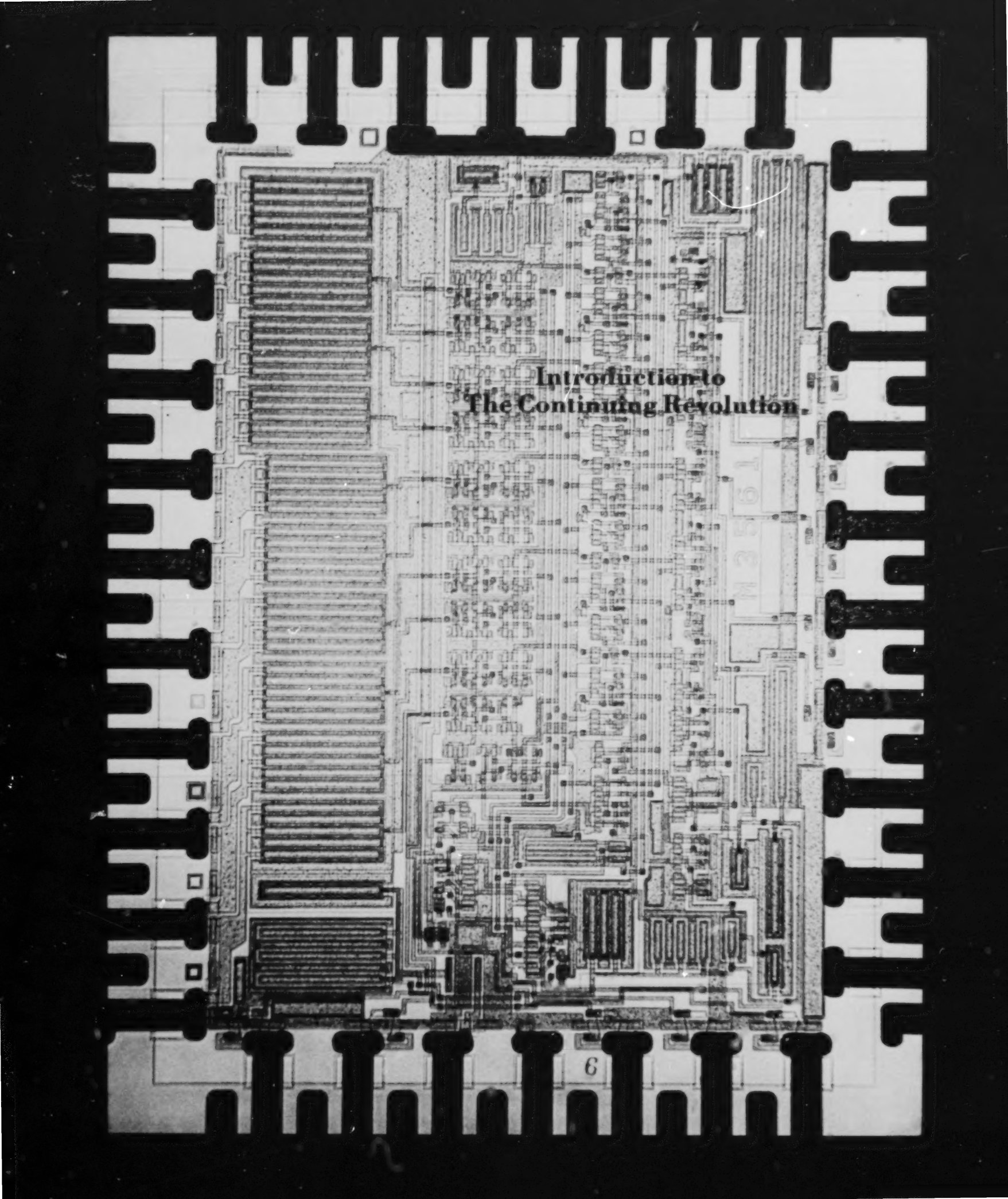
Physical Limits in Semiconductor Electronics: *R. W. Keyes* 197

Solid-State Electronics: Scientific Basis for Future Advances: *J. A. Giordmaine* 202

INDEX

Combined author and subject index 209

Courtesy Bell Laboratories →



Introduction to The Continuing Revolution

The Electronics Revolution

Philip H. Abelson and Allen L. Hammond

Earlier in the century, this country experienced a long era of sustained growth in many aspects. There was a steady increase in level of education, life expectancy, and standard of living. Growth of all kinds was welcomed, including industrial expansion and population increase. A feeling of progress, of achievement, of well-being was everywhere. As a corollary, a striving for excellence and the search for understanding were widely admired.

Today the mood of this country has turned pessimistic and negative. Those who are so inclined can find much evidence to support these views. Growth in the use of energy in the form of oil was suddenly curtailed in 1974. Growth in consumption of the kind seen in the 1950's and 1960's will not occur again. That part of the standard of living which is based on large-scale consumption of energy is not likely to improve during this century.

However, those who prefer optimism have reason for hope. Human ingenuity in solving problems is great. And native intelligence has been amplified enormously by the use of knowledge accumulated through research. An important product of research and a basis for hoping for a bright new future is the vitality of the electronics revolution. This revolution has been in progress for about 60 years. Lately its tempo has increased greatly. Until recently its importance was overshadowed by changes due to the large-scale expansion in the use of energy. But it promises to be more important, of more enduring consequence, than the earlier industrial revolution.

Some of the great changes brought about by the electronics revolution have gone comparatively unnoticed; at the start they were evolutionary rather than sudden and drastic. The telephone, which we take for granted, was invented 100 years ago. Nearly every decade since then the quality and scope of service have steadily improved, and the cost (measured in constant dollars) is now a tiny fraction of what it was 50 years ago.

Numerous applications of electronics gradually affected individuals and almost every component and activity of society. Radio was a marvelous toy and a source of wonderment when it was introduced 50 years ago. Now Americans listen to commercial radio an average of nearly 4 hours daily, and radio is accepted as practically a natural phenomenon. Television, which created a stir 25 years ago, is likewise commonplace.

During the last few years the impact of electronics on society has increased greatly. Examples are the rapid growth in popularity of citizens band radio, the worldwide use of the telephone, and the current astonishingly low prices at which hand-held calculators and electronic watches are being sold. Less evident to the individual but in total more important to society are other applications of electronics that affect nearly every sector of our economy.

This revolution, which is destined to have great long-term consequences, is quite different in nature from the industrial revolution. The industrial revolution was based on a profligate use of energy (mainly fossil fuels). Much of its technology was crude, with only a modest scientific or theoretical base. In large measure what the industrial revolution did was to make available and to employ large amounts of mechanical energy.

In contrast, the electronics revolution represents one of the greatest intellectual achievements of mankind. Its development has been the product of the most advanced science, technology, and management. In many applications electronics requires little energy. Indeed, one of the factors that guarantees enduring impact for the electronics revolution is that it is sparing of energy and materials.

With electronics one can control the disposition of large amounts of energy and force, but much in the way the brain is used in directing the action of muscles. In some aspects, electronics can be more subtle, more nimble, more dependable than the brain. In other applications, electronics serves as a great extender of human capabilities by rapidly carrying out routine but complex calculations, thus freeing the mind to make intuitive judgments and find shortcuts to new insights.

The industrial revolution, dependent on energy and materials, will be slowed and limited by the paucity of these necessary ingredients. The electronics revolution, fueled by intellectual achievements, is destined for long-continued growth as its knowledge base inevitably increases. Obviously, the current rapid rate of evolution of electronics cannot persist indefinitely, but significant change is likely to continue for a long time.

One of the factors contributing to this dynamism is that in laboratories devoted to extending the electronics revolution the use of powerful investigative tools based on electronics is speeding new developments. Moreover, the body of knowledge that is being accumulated in the natural sciences continues to grow, and its growth has been fostered by the new tools that electronics has provided. There are few laboratories devoted to studies on the frontiers of the natural sciences that are not dependent on one or many items of electronic-based equipment. Two examples indicate the extent of the impact. A human's speed of reaction is about one-fifth of a second. Measurements can now be made in times as short as 10^{-12} second. More important is the overall effect of electronic devices on quantitative determinations of many kinds. In some instances sensitivities have been increased by orders of magnitude while the times required for measurement have been diminished to a hundredth or less of those needed in earlier methods.

Dr. Abelson is the editor of *Science*. Dr. Hammond is a member of the Research News staff.

One of the factors favoring the development of electronics has been a comparatively high degree of social acceptance. There have been sporadic attacks on various electronic devices such as computers and there is continuing concern about privacy, but the intensity of criticism has diminished. In comparison to the number of objections raised to chemical products, to the environmental concerns associated with nuclear and fossil fuel energy, or to fears of recombinant DNA, objections to electronics have been few.

The average citizen is fearful of air pollution, for example, and is frustrated by a feeling that there is little an individual can do about it. In contrast, if a television program is offensive, it can be summarily dispensed with. Items that have recently become broadly available, such as the hand-held computer, electronic watch, and citizens band radio, enhance the public's feeling of participating in the benefits of electronics while not bringing with them discernible side effects. In future, electronics will provide many new tools useful to the general public.

The first four articles of this issue were designed to provide an overview of the past, present, and future of electronics. This is a large task, worthy of a shelf of books. Three major themes are treated in various ways in some or all of the articles: the tempo of the revolution, its magnitude, and the changing driving forces that have spurred it.

Until 1940, developments in electronics took place at a comparatively moderate pace. As was true with many scientific and technological matters, the pace quickened during World War II and was further maintained during the Cold War. Two major developments occurred independently during the late 1940's and later fused to give enormous impetus to electronics. One was the construction of programmable electronic computers. The second was the invention of the transistor. Subsequent developments in solid-state physics led to the present-day silicon chip with its large-scale integrated circuits. One such circuit in 1977 can contain more active elements than the most complex electronic equipment of 25 years ago.

After about 1960, when solid-state devices were incorporated in computers, there was a rapid development in the capabilities of computers and a steady reduction in the costs of calculations. An important effect of integrated circuits has been a reduction in the size and power requirements of electronic equipment that has made possible, for example, a Viking lander. Other advantages include reproducibility, maintainability, and reliability. Especially helpful is a sharp decrease in the need for making interconnections.

Large Scale Integration

The tempo of change has been impressive. In 1959 a chip that was commercially available contained one component of a circuit. By 1964 the number of components per chip had risen to 10, by 1970 to about 1000, and by 1976 to about 32,000. The cost per chip advanced only modestly. Thus, the cost per function has dropped drastically. It is this great change in the cost-effectiveness ratio that has made possible inexpensive hand calculators and related microprocessors and minicomputers. One of the key individuals who has been pushing the development of large-scale integrated circuits is Robert Noyce. His article in this issue discusses the development of these devices with special emphasis on what is yet to come. He shows that further advances can be expected. Theoretical considerations show that physical limits have not been approached. He is so bold as to state that "if the present rate of increase of complexity were to continue, integrated circuits with 10^8 elements would be available in 20 years."

All of us have seen examples of the sudden termination of exponential growth, so perhaps Noyce's figure will never be attained. But substantial advances toward his goal are already in progress. He seems justified in the view that "the potential for developing inexpensive processing power is truly awesome." He projects that with low-cost processing many new tasks will be undertaken that are uneconomical today.

Another way of glimpsing the tempo and magnitude of the electronics revolution is to focus on what has been happening in computers. In the early 1950's, almost all computers were owned by or devoted to tasks of the federal government. Computers were procured for use in such applications as defense and nuclear reactor design. By the mid-1950's there were about 1000 large-scale computers, and the tendency was toward increasing computational power. By the mid-1960's, there were 30,000 computers, and the generally accepted view was that costs of computation decreased with size; that is, the larger the better. At the end of 1976 there were about 220,000 computers in the United States. Of these, 40 percent were medium or large computers; the remainder were minicomputers which are small and by definition cost less than $50,000. At the same time, there were 750,000 of the microprocessors that form the heart of microcomputers. Ruth Davis estimates that by 1980 the number of minicomputers will reach 750,000, while the number of microprocessors will increase to more than 10 million.

As the number of computers in service grew, the uses and the organizations involved broadened. The Davis article summarizes the current distribution of ownership of conventional computers. The percentages in the major categories are: manufacturing industry, 31; miscellaneous business, 13.3; financial institutions, 13.4; wholesale and retail trade, 13.1; educational institutions, 5.7; state and local government, 5.7; and federal government, 3.4. There is further scattering of ownership throughout virtually every kind of organized activity. Thus it may seem that an enormous shift in the nature of the market for large computers has occurred. Beyond that is the larger market for minicomputers and the much larger mass demand for microprocessors.

Coincident with the expansion in the number of computers has been an increase in the number of computer professionals. During the past 20 years the total number of analysts, designers, programmers, and operators has increased from 100,000 to 2.5 million. The number of students having some degree of familiarity with computers is much greater. This reservoir of people familiar with applications of computers is certain to facilitate additional applications of electronics. The emergence of computer hobby shops is bringing additional enthusiasts and imagination into the field. One group that is likely to make substantial contributions is the working scientists in the natural sciences. Often their progress and ability to tackle problems are limited by their equipment. Having experienced the advantages of incorporation of microprocessors in measuring devices, they will be looking for novel kinds of electronic sensors that can be coupled with the current data processors.

Because many of the new major applications involve various kinds of computers, one might have the impression that the electronics revolution and computers are synonymous. It is easy to lose sight of the importance of the noncomputer aspects of electronics. Key to many applications are the transducers or sensors. For example, computers would have a limited role in process control if electronic devices for sensing temperature, pres-

sure, and concentrations of components were not available.

The potential applications of electronic technologies are so numerous and so provocative as to give free rein to futurologists and science fiction writers. The domestic robot, the wired city, the global electronic village—none of these can be dismissed as being beyond the bounds of technical feasibility. But it is not necessary to look so far afield to see how pervasive the impact of electronics is, how many areas of human endeavor and how large a portion of the country's economic activity may be substantially altered. Indeed, it is probable that reality will outstrip fiction in the rate of introduction of new and often unexpected applications of electronics in coming years. Witness, for example, the incredible growth in popularity of citizens band radio. It is clear that the capability of some electronic devices, particularly microprocessor circuits and memory units on single silicon chips, is developing more rapidly than applications can be conceived of and introduced.

The markets for kitchen appliances, office equipment, and leisure games, to mention just a few, are ready to be revolutionized or at least substantially modified by the addition of logic and memory to yield "smart stoves" and similar products, the first of which are now available.

The driving force behind many of the commercial applications is the extremely low prices for sophisticated electronic circuits, which in turn derives from mass production. The key innovation allowing such large markets is the microprocessor, a general-purpose logical unit that can be programmed to perform an unlimited number of tasks, thus eliminating the necessity of designing new circuitry for each new application. Among other applications, microprocessors are making it possible to extend computer control to mechanical and electrical equipment of every description, from consumer appliances and automobile engines to milling machines and industrial boilers. In the past, automation of manufacturing and process control has moved slowly because of fear of dependence on a central computer and the cost of the controlling units. The first process control computers introduced in the late 1950's, for example, cost about $300,000; minicomputers reduced this to less than $100,000 by the late 1960's; now microprocessor controllers are available for $3000, cheap enough to automate control and data collection for even small process steps. What seems to be evolving is a linked, hierarchical arrangement in which microprocessors are used to control individual pieces of equipment; minicomputers collect and process management information from the microprocessors for an entire factory; and large central computers use the resulting data in compiling corporate financial reports.

But the impact will not be confined merely to consumer products or isolated devices. The application of electronics is already having a pervasive effect on the entire economy and on our way of life, one that promises to intensify in coming years. Consider only a few of the areas treated in this issue—medicine, education, national defense, banking and retail sales, postal and other communications, and the research process itself.

The practice of medicine, for example, has already begun to change in such areas as the handling of patient records, billing and other administrative chores, and computer-controlled examinations in response to conventional data-processing equipment. Even more fundamental extensions of the physician's skill are resulting from the application of compact integrated circuitry to diagnostic and monitoring equipment. The potential of medical electronics is indicated by the unprecedented demand for tomographic x-ray scanning equipment, which by computer processing and synthesis is able to distinguish different tissues with a sensitivity 50 times that of ordinary x-ray techniques; hundreds of these new diagnostic tools have been ordered. Another new and noninvasive diagnostic approach, the use of acoustic waves in such devices as ultrasound cameras, is also beginning to be widely applied; here the key role of electronics is to translate the acoustic information into visual and analytical data. Perhaps the most striking illustration of the unique power of electronic circuits in medicine is their potential use as prostheses to supplement or replace damaged neural tissue, a circumstance that is only possible because modern circuits now approach the size, power consumption, and logical capability of the natural tissue. The cardiac pacemaker is an early example of such a prosthesis, and, as R. L. White and J. D. Meindl point out, the development of far more complicated devices such as an implantable electronic ear for the deaf is well under way.

Potentially, electronics and electronic media could have an important impact on education, as almost anyone who has observed children watching Sesame Street could confirm. Despite a few such successes, however, there seems to be general agreement that television and computer-assisted instruction have not yet lived up to that potential. But educational innovators have not yet given up, and one new approach is proposed by J. F. Gibbons in this issue. There is another sense, however, in which electronics is certain to affect education for better or for worse, and that concerns the prospective flood of inexpensive electronic devices of which the hand-held calculator is only the first. Calculators have substantially altered the character of the traditional "problem sets" in science and engineering courses at the university level, they are becoming common in high school courses, and they are already creeping into use in primary schools. Some parents and educators are trying to stem this growing tide on the grounds that it will only add to the reasons "why Johnny can't add." Others see the trend as inevitable and point out that how computation is performed is irrelevant, what really counts is whether the students learn the underlying concepts, and in this respect the impact of the calculator is still uncertain. In any case, the ubiquity of the calculator seems to guarantee that electronic arithmetic will become the language of the real "new math," and these developments suggest the potential of the more elaborate calculator-based games and educational devices that are beginning to appear either alone or as attachments to the home television set.

It is difficult to imagine a modern military force without heavy dependence on electronics. Aircraft instrumentation, missile guidance systems, radar and other surveillance sensors, tactical computers—all depend on electronic components. But electronics plays more than a passive role in military systems; in recent years, advances in electronics have been perhaps the most important factor guiding the evolution of new weapons and new strategies. One example is the emergence of "smart bombs" and other unpiloted weapons, which can evaluate guidance information to track themselves to target or make use of preprogrammed instructions to maneuver evasively. Carried to its logical conclusion, this trend might eliminate the need for many manned aircraft and is at the core of the current debates over the B-1 bomber and the cruise missile. A second example is the NAVSTAR satellite system, for which prototypes are now being tested. These navigational satellites are designed to allow any military vehicle carrying an inexpensive receiver and computer to instantaneously determine its position anywhere in the world with an accuracy of better than 10 meters in horizontal and vertical coordinates; civilian aircraft and ships will also be

able to use these satellites, but with somewhat less accuracy. This phenomenal accuracy is expected by many defense analysts to revolutionize navigation, weapons targeting, battlefield management, and other aspects of warfare; this is especially true for fixed targets, since the coordinates of any such target can be readily determined. The system also may supplant many of the commercial navigation systems now in use.

In 20 years, the role of computers in research has been transformed from what W. O. Baker and his colleagues describe as "a minor annex of mathematics research" to a major and often dominant role characterized by the proliferation of dedicated minicomputers and time-shared terminals in most research institutions. They assert that computers have transformed the research process from conceptualization to experimentation to publication. Baker should know, since his organization has been in the forefront of actually putting computers to work in research—Bell Laboratories now have an average of one dedicated minicomputer and five interactive terminals for every 15 professional staff members. Electronics has also transformed other instruments of scientific research, from the electron microscope to vidicon astronomical cameras, and the process is accelerating as more and more "smart instruments" are designed around microprocessors.

Nowhere is the potential impact of electronics greater than in the banking industry and the Postal Service, both of which face the prospect of converting from moving pieces of paper around to using electronic transfers for at least part of their business. These changes will certainly not come overnight and will raise a host of social problems; how does one protect against theft when most retail transactions are done electronically rather than by check, or guarantee privacy if much first-class mail travels by wire? But electronic transfers of money, of messages, and of documents are already established features of our society. The U.S. government, for example, makes some 5 million Social Security payments each month by sending banks magnetic tapes that the bank computers will use to credit depositors' accounts directly. Preauthorized, nonpaper payments are estimated to account for 10 percent of bank transactions in some areas of the country. Point-of-sale electronic terminals are now becoming common in retail stores, although their role is presently restricted to credit verification and inventory control, not direct transfer of funds from the customer's account to the store's. But with 26 billion checks a year passing through the banking system and the likelihood that the volume will double by 1985, there is ample incentive for banks to move toward electronic transfer and checkless banking. As F. E. Balderston and his colleagues point out, such a move will change the boundaries between the retail trade and the banking system, possibly resulting in more decentralized but far more complex financial networks.

Checkless banking, if and when it does occur, will intensify the economic pressures on the Postal Service, since nearly 40 percent of the mail consists of checks and other financial transactions. Diversion of this mail will reduce revenue but will not noticeably lower costs. But that is not the only threat. As complaints about lost or delayed mail increase, many large businesses are looking toward electronic mail systems. One prototype of such a system is the Department of Defense's ARPA computer network, which is routinely used by researchers all across the country to exchange messages and information. New optical scanning and electronic printing techniques are being developed by many companies that would allow users to transmit documents or whole pages of text. The Postal Service may be forced to embrace electronic mail or face the future as an obsolete, increasingly expensive system serving fewer and fewer people.

Computers and Communications

Postal communications are not the only form of communications facing new challenges. Not long ago radio and marine cable telephone circuits were the principal means for rapid intercontinental communications. Now an international satellite communication system is well established and carrying a growing volume of traffic. Domestic satellite systems are just getting under way in the United States, but they seem certain to expand the options for voice, television, and digital data communications. The transmission of digital information is the most rapidly growing area of electronic communications and reflects the increasing need for computers and other intelligent machines to "talk" with each other. Indeed, as D. Farber and P. Baran point out, computer and communications technologies have become so similar and intertwined that they are difficult to distinguish. More and more communications are transmitted in digital form, even within the telephone system. And more and more information in the communications systems is processed both before and after transmission; voice signals are compacted and compressed to put more calls on a channel, for example, or the output of the intelligent terminal on one end of the phone line becomes the input for a computer or a display device on the other end. Distributed processing—essentially networks of small and medium-sized computers connected by communications links—is clearly going to be one of the major forms in which computers are used.

These developments clearly pose a major problem for those, such as government regulators, who must decide where communications—a regulated activity—ends and where the unregulated computer market begins. Elsewhere in this issue M. R. Irwin and S. C. Johnson summarize the public policy issues involved in this intensifying conflict, and point out that among the principal contenders are some of the giants of American industry, AT&T and IBM. It promises to be a multibillion-dollar fight and one of the thorniest technological policy problems the government must face in the coming decade.

The evolution of computers to the point where communications is a major part of their activity is also reflected in other changes. The traditional use of computers as calculating engines for numerical work is rapidly being replaced by a new principal role, that of managing, storing, retrieving, and distributing information. A search for new computer architectures that better reflect this new role is under way and includes experiments with such things as augmented sets of instructions for the computer's own control program, specialized subcomputers or processors to manage databases, and multiprocessor machines. The generation of computer programs—all too often a bottleneck to effective use of computers, as H. D. Mills points out—is increasingly being put on a firm mathematical basis. The way information is stored in computer systems is changing too, as researchers look for more efficient search routines, new methods of combining memory devices, and new computer languages adapted for information processing. S. E. Madnick, describing this evolution, believes that what is emerging is the goal of an information utility that can serve many users for many purposes. Prototype experiments with information utilities are actually under way in Britain, where the British Broadcasting Corporation is testing a small decoder attached to television sets that can, on demand, deliver current information on a variety of subjects; the British Post Office, which runs the telephone network, is also experimenting

with a telephone-based information utility that would combine the functions of a daily newspaper with the resources of a library.

We are increasingly an information-based society. Economically, for example, information industries ranging from broadcast television to book publishing to computer services contribute a large part of the U.S. gross national product and employ nearly half of the work force. The information sector of the economy is also among the more rapidly growing. Moreover, information is a resource that greatly enhances individual capabilities and opportunities and is not depleted by use. H. A. Simon describes the development of the ability to process and manipulate information on a large scale as having a significance, on the scale of human evolution, equivalent to the development of written language or the invention of the printed book. In any case, it is clear that the information revolution will accelerate as more persons acquire their own computers and as these computers are able to make use of larger and larger information resources.

The conventional economic wisdom is that the expanding opportunities in information-related activities will more than offset jobs lost to more productive electronic equipment. Recent experience in the electronic manufacturing industries would seem to bear this out, although substantial layoffs in some companies have been avoided only by a commitment to large retraining programs. In the coming decade automated electronic equipment is likely to make inroads into the service sectors of the economy as well. The nature of secretarial and other office support jobs may change, for example, as may that of mail clerks and bank tellers. Whether any of these changes will result in displacing large numbers of people from these traditionally labor-intensive occupations is not clear, but the process of change is certain to be uncomfortable for the individual whose job is involved and maybe for society as a whole. Consider the impact of converting the Postal Service, one of the largest employers of unskilled and semiskilled labor in the country, to an electronic mail system.

Despite such problems the electronics revolution is not likely to slow down anytime soon, if only because the research base is broad and vigorous and is already producing a host of new ideas and new concepts that are certain to be translated into new products and services in coming years (Fig. 1). One trend that can be identified is the incorporation of magnetic, acoustical, and optical phenomena into electronic devices, giving rise to a host of new effects that can be put to use. Examples include magnetic bubbles in memory devices, surface acoustic-wave filters in signal processing equipment, and optical fibers in communications. Particularly fascinating in this regard is the push toward optical communications.

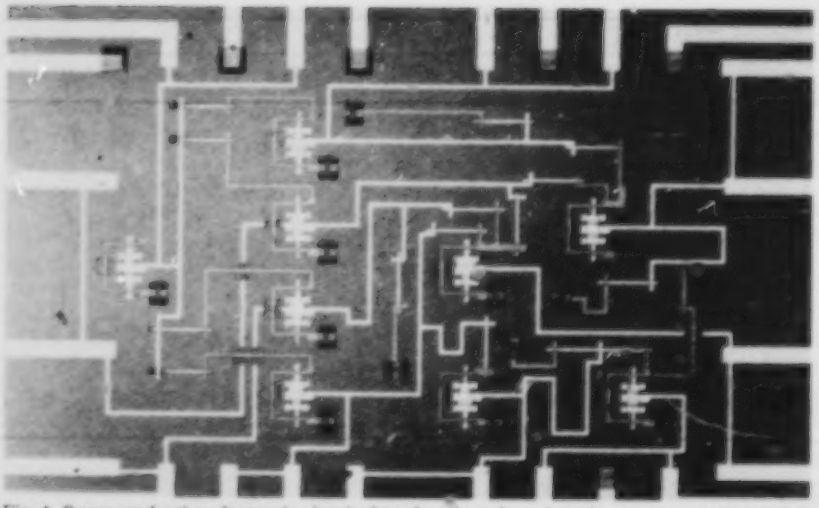


Fig. 1. Superconducting electronic circuits based on Josephson junctions are one or two orders of magnitude faster than conventional transistor circuits and consume a thousand times less power. This technology, which has yet to be incorporated in computers, may initiate a new era of extremely rapid processing equipment. The circuit shown here contains 18 Josephson junctions, operates at 4°K, and has an add time of less than 1 nanosecond. [Source: J. Magerlein, T. A. Fulton, and L. N. Dunkleberger, Bell Laboratories]

The concept of transmitting information on a light wave dates back at least to Alexander Graham Bell, who in the 1870's demonstrated a wireless telephone based on light that could transmit sound for more than a kilometer. A hundred years later, optical communications is on the verge of becoming a reality. The principal advantage of communicating with light waves is their high frequency, compared with radio waves, which gives them a superior capacity to transmit information. A single optical fiber, for example, can carry hundreds of times as many bits of information per second as a copper wire. The Bell system is already experimenting with an optical link for interstation connections in areas with a high volume of calls, and others are looking at applications ranging from computers to military vehicles. As large-scale production of fibers and other components gets under way, cost reductions as dramatic as those in electronic calculators are expected.

In its modern configuration, optical communications relies on such components as lasers or light-emitting diodes for light sources and glass fibers for the transmission medium. It is, in fact, the dramatic improvement in these components that has brought the technology to the verge of utility. S. E. Miller reports, for example, that optical fibers have been produced with losses less than 1 decibel per kilometer—less than the loss in light passing through a single windowpane. But, just as solid-state technology progressed from individual devices to integrated circuits, integrated optical circuits incorporating lasers, amplifiers, and detectors on a single chip are already being developed. Many semiconductor materials, it turns out, are optically as well as electrically active, thus permitting the intimate interplay of both kinds of circuits. This activity is leading to an era in which the electron, long the workhorse of the electronics revolution, will be supplemented by the even greater potential of the photon.

The range of phenomena and the indications of still-to-be-exploited potential to be found in electronics research are convincing evidence that we have not yet seen the limits of what is possible. Still less are most of us and most of our institutions prepared to decide what we should do with our new capabilities or even how to cope with the speed at which electronics technology is changing the ground rules under which we operate. In business the price of being unprepared is often high, as many white goods manufacturers found out when one of their competitors introduced an electronically controlled microwave oven that has rapidly become the best-selling product in its field. For governments and individuals alike the stakes are arguably lower at present, but the continuing electronics revolution promises to be so pervasive as to compel the attention of even the most unobservant.

Electronics: Past, Present, and Future

We are so engulfed by electronics that we can scarcely gauge the impact of the art on society.

John R. Pierce

What *is* electronics? Once we associated electronics with vacuum tubes, but vacuum tubes are almost obsolete. Perhaps electronics is semiconductor devices. But then, what of magnetic cores and bubbles and liquid crystals? I think that electronics has really come to mean all electrical devices for communication, information processing, and control. This would include pre-vacuum-tube devices such as the electric telegraph, which replaced such light-wave communication as semaphore telegraphs, signal flags, and heliographs. It would include the telephone.

According to this definition, electronics was born nearly 150 years ago with the first electric telegraphs. One hundred years ago we had transoceanic telegraphy and the first crude telephony. Seventy-five years ago the telephone system had had a substantial impact on life but the continent had not yet been spanned. Fifty years ago we had transcontinental telephony, a good start on automatic switching, a crude start at transoceanic wireless telephony, and the beginnings of radio broadcast. Twenty-five years ago television had begun, we were on the verge of transoceanic telephony by submarine cable, and we had the first large and effective computers. Fifteen years ago we had the beginning of satellite communication and computers were out of the laboratory and into industry. Ten years ago computers were shrinking in size and growing in speed. Five years ago we were on the eve of what few expected—a pocket calculator more powerful than the first computer. And today? Today we are so immersed in, engulfed and almost buffeted by electronics that we can scarcely know the nature, extent, or impact of that art.

The author is professor of engineering at California Institute of Technology, Pasadena 91125.

Except for the computer, the things that I have mentioned show the impact of electronics on one field—communication. Largely, but not entirely (as x-rays remind us), electronics in other fields grew out of the electronics of communication. Today, communication, expanding as it still does, is only a part of electronics.

I think it is important to give some account of the widespread and astonishingly various triumphs of electronics. We frequently read of the triumphs of the computer in chess, or in recognizing patterns, or in transcribing speech, or in reducing profoundly the bandwidth required to transmit television, or in proving theorems, or in controlling traffic, or in providing information instantly and cheaply.

Electronics is, indeed, wonderful. But how wonderful?

I have made my own short list of things which I believe illustrates the wonders of electronics. It is certainly incomplete. It may be in error. The list includes three sorts of things: Those that would be impossible without the speed and accuracy of electronics (Table 1), those that would be exceedingly difficult or imperfect without electronics (Table 2), and those that are greatly aided or improved by electronics (Table 3).

Some of the entries may seem rather surprising. Surely, effective nuclear energy would have been impossible without the part electronics has played in discovery, design, and control. Archeology and electronics—what of carbon dating? In a recent report it was said that the four-color theorem has finally been proved. The computer played a crucial part. Physicists and others use computers to manipulate polynomials so complex that a trustworthy answer is completely beyond pen and paper.

In connection with medicine, biology, and chemistry we can recall nuclear resonance, spectroscopy, and measurements involving radioactive isotopes. Production of complex sounds is included because computers have been used to produce highly organized sounds of musical intent, sounds which, whatever their merit, could have been created in no other way.

In industry, numerically controlled machine tools are used in shops, large and small, here and abroad. Electronics has become an essential part of automation and process control. Planning and accounting, from the point of sale of a product to the annual report, have changed greatly as a result of electronics, as has banking.

It seems appropriate to ask whether, by considering the multifarious aspects of electronics and the world in which electronics plays a part, we can draw any general conclusions. So far I have tried to indicate the deep penetration and broad influence that electronics has had in our technology and our lives. In doing this it is difficult to distinguish fact from fiction and lasting impact from clever trickery.

There is quite another aspect of electronics and our experience with it. This is the insight that electronics and its use have afforded us on how science works in society. Electronics provides us a broad range of interactions of science with society that is at once heartening and illuminating.

One matter on which electronics casts light is the possibility of improving productivity in the service sector. It is a commonplace that increased productivity makes goods cheaper before inflation, while services do not become more productive and rise in cost in any terms. Electronics has provided services in which both output and labor can be measured quantitatively.

In telephony, the measure of output or use is the number of telephone calls per year and the measure of labor is the total number of employees. Over the past 50 years, the number of calls per employee has risen at a rate of about 2.5 percent a year, as shown in Fig. 1. This has been attained through automation of telephone switching and automation or partial automation of many testing, maintenance, operator, and billing functions.

In a day when opera, concerts, and the stage suffer from a constant audience size and ever-rising salaries, electronics has preserved the economic viability of music, drama, and entertainment. With the aid of electronic amplification, a small rock group can be heard "live" by tens of thousands. Records bring one

Table 1. Things that would be impossible without electronics.

Telegraphy
Telephony
Radio
Talking pictures
Public address systems
Television
Data communication
Radar
All automatic service and vending (except slot) machines
Management of airports and air traffic
Controlling moment by moment the total power used in industry
Radar and radio astronomy
A good deal of work in physics, chemistry, biology, psychology, statistics, mathematics, archeology, seismology, optical astronomy
Some medical diagnosis and treatment
Nuclear energy and nuclear weapons
Ballistic and guided missiles
Production of highly organized complex sounds
A machine to play chess

Table 2. Things that would be very difficult without electronics.

Calculators
Automation and process control
Guidance and navigation
Phonographs
Numerically controlled tools and processes
Collision avoidance
Large-scale testing
Seismography
Analysis of complex structures
Complex machining and fabrication
Much medical testing and treatment

Table 3. Things that are substantially aided or improved by electronics.

Surveying
Banking
Retail sales
Typing, editing, publishing
Operation of automobile engines
Repressive government

performance to millions. Through television, a performance can reach a first audience of tens of millions, and reruns go on forever.

Electronics shows that productivity can increase in the service sector.

A demonstrated increase of productivity in the service sector is something that anyone can understand and applaud. A demonstration that science, industry, and the public interest can exist symbiotically is something that scientists should particularly applaud, because this demonstration helps to further their own interests, spiritual and temporal. Electronics has clearly inspired and supported scientists.

Solid-state physics was a sleepy subject before the invention of the transistor; it has become a major field of university research. All sorts of discoveries of great intrinsic and practical importance have followed: the elucidation of the band structures of various materials; the discovery and understanding of surface states; the devising, understanding, and exploitation of new compounds; and, most recently, the discovery of new states or phases in semiconductors illuminated by laser light—first a condensed phase of electrons and holes, and now a new metallic phase which is so strongly conducting that it can be used to switch a signal in 10^{-12} second.

Ranging a little further afield, there has been a similar growth of understanding in magnetics, in superconducting phenomena, including the exploration and use of the Josephson effect, and in the understanding of the nature of materials, including superstrength crystals which have no dislocations and spinodal copper alloys which derive their strength through a sophisticated control of dislocations.

Particularly worthy of mention is the production of transparent silica-based fibers a hundred or more times as transparent as the best glass. An attenuation of around 4 decibels per kilometer at 0.8 micrometer was attained by Corning in 1972; around 2 db/km at 0.85 μm by Bell Laboratories in 1974, and about 0.4 db/km at 1.2 μm by Nippon Telephone and Telegraph in 1976. Such fibers, used for communication by light waves, promise to be the most revolutionary advance in this field since the transistor and integrated circuits.

We should note that the very invention and production of integrated circuits has led to a new examination of materials, epitaxial growth, diffusion, the use of ion beams for both analysis and implantation, and to fundamental studies of the limitations of size, energy, and speed in the operation of semiconducting devices.

We should also note that the exploitation of electronics led Shannon to formulate in 1948 a penetrating theory of information which has had considerable consequences for mathematics as well as for communication.

Advances in electronics have produced social phenomena which all who respect both science and society desire—a body of scientific work which is shared among many institutions, and shared quite freely, which is of great scientific merit in itself, which has a clear value to and a beneficent influence on man's society, and which is of direct use to individual men. Who would want to give up his pocket calculator, let alone his telephone or TV?

Why has this interrelation among science as discovery, industry as provider, and government as a supporter of research in universities and elsewhere and as an initial market for such products as computers and integrated circuits been so agreeable and productive in electronics? Partly, I think, it has been a matter of selection and concentration.

The original thinker and experimenter can do only so much by himself and with limited equipment. Sometimes, fields of work have been pursued by many because of other people's new ideas and because the field has high prestige. Quantum mechanics and high-energy physics have progressed through deep insights which inspired workers and, of late, through massive government support.

Sometimes fields have attracted able minds into association and cooperative endeavors through government support toward goals quite other than giving every man something for which he will pay directly after taxes. The atom bomb and

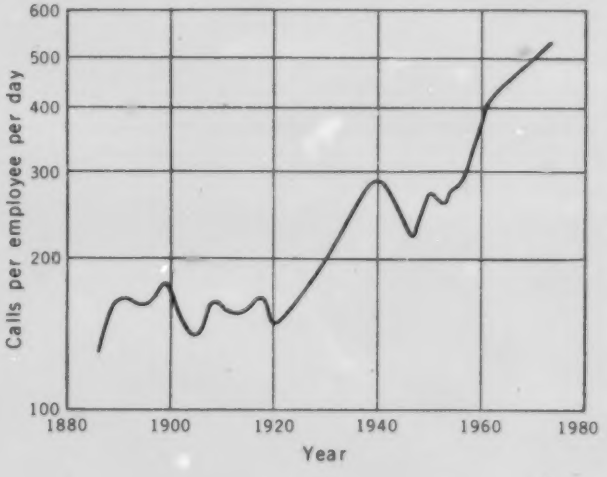


Fig. 1. Electronics can increase productivity in the service sector. The number of calls per telephone employee has increased at a rate of about 2.5 percent per year since 1920.

the space program are examples of successful cooperative-scientific endeavors toward selected goals, goals much narrower than those of electronics.

Electronics attracted ingenious men to a reasonably defined area of work, it offered both scientific and commercial success, and, through this success, it provided support for the scientific work it had engendered. A good deal of wise support has come from the federal government, and taxes derived from electronics have flowed to the government to cover the support of electronics and other expenses. Further, our balance of trade has been saved from complete disaster only through the export of agricultural and high-technology products, and electronics has played a considerable part in providing high-technology exports.

Thus, electronics has created within our society a productive association among scientists and industries which has been profitable to both, and to consumers as well. In so doing it has answered the question, What sort of industry is better, large or small? The answer is both, and a little more.

Without small industries, many things would not have happened as quickly as they did. This includes transistor radios, hearing aids, and really good oscilloscopes. There are all sorts of other examples. Small companies move fast. They have to be quick and successful with a particular product in order to survive.

Successful small companies do more than supply the user with an attractive product at an early date. Very successful small electronics companies become larger and more diversified companies. Beckman Instruments, Varian, Techtronics, and Hewlett-Packard are a few examples among many. These firms supply a host of users with a variety of complex and sophisticated electronic devices on which science and progress have come to depend.

Important as they are, small and medium-sized companies draw the science on which they thrive, and sometimes the invention, from outside sources. They devise, perfect, and market rather than discover. The sort of inspired research of doubtful outcome which brings revolutionary advances is done elsewhere. Sometimes it is done in universities, but much is not. Academic activity seems not to span all socially important aspects of nature, or perhaps not to pursue them deeply enough. Much original discovery and the invention based on it has come from large industrial laboratories.

Industrial laboratories can carry out long-range work only if they are attached to large, stable industries which support them through thick and thin. Valuable discoveries are made on the average, not regularly at the end of each fiscal year. A small or even a medium-sized laboratory cannot afford even a productive long-range program of discovery. Enduring as government laboratories have been in pursuing many worthwhile endeavors, by and large the process of significant discovery has eluded them.

Thus, tangentially, the success of electronics has answered the question, should business be large or small? Success depends on both kinds. Small business has been quick to improve and exploit. It was no accident that the first successful microprocessor did not come from a giant in the computer industry. But large business has provided much of the ground from which discovery has sprung.

Electronics has interacted with all areas of society, including government in its regulatory as well as its supportive function. In this encounter, the stifling effects of regulation, even in an area as scientifically powerful and socially useful as electronics, have been made painfully apparent.

A high quality of telephony with moving vehicles is technologically possible at a price which would probably be attractive to a large number of users. What we have is a pittance of poor service and citizen's band radio. The latter is wonderful as a sport and influential on our language. We quickly recognize the true double-talk in which each statement is repeated as a consequence of trying to make ourselves understood over CB. Neither present mobile telephone service nor CB is an extension of what we should ask of telephony.

Our lack of a mobile service which is a compatible extension of telephony is entirely a consequence of government action, or inaction, in assigning adequate frequencies to someone who will provide adequate service.

The first successful active communication satellite (Telstar) was launched in 1962 and it carried live TV across the ocean. The first commercial satellite transmission was carried by Intelsat I (an outgrowth of the Syncom of 1963) in 1965. While international satellite traffic has increased rapidly in the 11 years since then, the first American domestic satellite was not launched until 1975. While six American domestic satellites are now in orbit, their use is limited, partly because of regulation of use, and partly because regulation has prevented the commercial use of the higher frequencies that are more suitable for domestic purposes.

Thus, government regulation has retarded the use of satellites for domestic communication. We may note in passing that government support of propagation measurements at higher frequencies was late and niggardly. For years the only measurements available were those made by industry, using the sun's microwave radiation and the microwave temperature of the sky as tools. Even detailed knowledge of the pattern of rainfall was obtained by industry rather than through government studies.

We may also note in passing that successful as satellites for overseas communication are, their vulnerability poses a threat to us which most do not appreciate, and the way in which our government decreed that satellite service be provided makes it seemingly impossible to do anything about this vulnerability.

Many have hailed cable TV and pay TV via cable as the best hope for mass communications in the future. They see many channels and new sources of revenue as an economical, commercially viable way for providing programs to segments of the public as well as to the whole TV audience assembled. Surely, cable TV and pay TV can give us local as well as regional and national news and advertising, and programs aimed at groups with special interests as well as programs aimed at the least common denominator. Or can they? Not as long as the regulators procrastinate and make inordinate demands for more complicated and higher quality service.

Not all important aspects of electronics involve all of society. Many problems, and particularly problems of a systems nature, arise from the very nature and power of electronics itself.

To me, the fact that I can pick up a telephone, dial a person in a distant city, talk, terminate the call and find it itemized with the number called on my telephone bill, all without human intervention, is more amazing than that a computer can be programmed to play a fairly good game of chess. The direct-dial, long-distance, automatically billed call is wonderful because of the complexity of the total system involved, which among other things, seeks an alternate route if the primary route is not available, and which functions even when some routes are destroyed. It is particularly wonderful because all sorts of different equipment of many vintages somehow operate together to give a standard service, and because this service is so reliable, working because of and in the face of continual maintenance.

Further, if we were to look inside the system, we would find a surprising amount of automated planning, testing, and maintenance.

Making a computer play chess is something that a single clever man can understand and do. Making a huge, complex system like the telephone network is something that no single man can do, nor can one man understand the telephone network in complete detail. The whole network operates because various pieces have been made to operate reliably in a prescribed manner, and a way has been found to connect together pieces whose functions, as seen from outside the black box, are known and orderly.

The sort of large, multiprocessor system that the telephone network is leads us to problems of complexity in digital systems such as computers. I can remember the days (about a decade and a half ago) when it was axiomatic that the best computer was the biggest computer. Grosch's law (now repealed, I believe) told us that the cost of computation decreased inalterably as the size of the computer was increased.

Computerniks met the challenge of complex systems head on. Operating systems and languages became capable and complex almost without limit. Computer manuals grew from slim volumes to 5-foot shelves of books. The computer was capable (between crashes) of doing anything for everyone, and everyone was urged to put his problem on one big machine.

When even the most primitive small transistorized computers came along, battered users fled the large systems and put their problems on their machines. Small computers equipped, adapted, and programmed to carry out a particular range of functions very well are now commonplace in laboratories, industries, businesses, and military systems. And, a good deal of computing that would once have been done on the computer is now done on hand calculators.

However, a small computer will not do everything. Sometimes the owner of a small computer will want to use the facilities of a large central computer to plot data, or to draw pictures, or to print out elegant text, or to compile programs. He may wish to use a central file system in editing programs or text. He may want to use a central computer to transfer data to some other computer.

Thus, a small computer may become part of a computing system. Like a local telephone office, one computer can do some things by itself. In order to do other things it must call on other parts of the system.

The parts of the system it calls on may not be computers. They may be plotters and printers. Other components may be smart instruments which take data or even run complete tests. Others may be efficient special processors. In the commercial area, point-of-sale and other terminals must communicate with host computers.

The problems of digital electronics are great. They involve more than the search for algorithms which will enable a computer to play chess, or retrieve information, or speak, or listen and type out. They involve the efficient organization of various limited or special purpose digital equipment into a system which will accomplish some overall result in a useful and effective manner. In many systems, such as those used for credit card verification or electronic transfer of funds, equipment in various geographical locations is of different manufacture. The inside of one piece of equipment is unknown to another piece of equipment or to its designer. Only the overall function provided and protocols of communication are common among various installations.

Data communication systems themselves must somehow conform to the requirements of the equipment they interconnect. Two extreme solutions are to provide translation among all possible protocols and to buy equipment from one manufacturer only. An intermediate solution will have to be found.

Whatever aspect of electronics we consider, we find it difficult to distinguish among dreams of the future, misconceptions of the present, and actual accomplishment. Electronics seems too extensive to evaluate.

Yet is is clear that electronics is a necessary ingredient, not only of the communication, out of which it grew, and computation which is a chief present manifestation, but of transportation, nuclear energy, and a great deal of various commercial activity and in scientific research of all sorts.

When we consider electronics in detail, we find that progress has been greatest and quickest in providing ever faster and cheaper memory and logic. The speed of electronics assures that electronics must surpass human capability for many tasks.

Despite problems of complexity, computer scientists have produced algorithms which solve very difficult problems, once those problems have been well posed. Thus, computers far exceed human capabilities, not only in numerical analysis, but in guidance and control; in the manipulation of polynomials; in keeping books and in making out income taxes (once the accountant has made certain crucial decisions).

Indeed, computer users have been ingenious in finding problem areas in which computers excel man by far. Computerniks also have ingeniously found a number of problems at which computers are inferior to man, including understanding speech, recognizing complex patterns (people, for instance), and playing chess.

However much progress we have made, there are still many unsolved problems in electronics. One of these is how to assemble digital equipment of many sorts from many manufacturers into large systems which will perform complicated tasks economically and very reliably.

Another problem is how to produce small, cheap, reliable interfaces to link cheap, fast logic and memory to tasks in the world around us. Such interfaces include displays, plotters, printers, keyboards, communication gear, sensors of all sorts and natures, simple and complicated, and all sorts of activators.

Surely electronics will progress as rapidly and as profitably and as complexly in the future as in the past. We will always think that we know where it will be in the future. We will continue to have misconceptions concerning the state of electronics at any particular time.

Evolution of Computers and Computing

Ruth M. Davis

The evolution of computing and computers is particularly attractive to depict since most adults of today have lived through it in its entirety. Computers, as I will be describing them, have just (in 1976) celebrated their formal 25th anniversary. As 25-year-old computing devices, electronic, digital, stored program computers are very young.

Indeed, any comprehensive chronology of computers (*1*) would always trace the origin of computing devices to the counting table around 400 B.C. and the abacus used in China as early as 500 B.C. This chronology would also include the history of machines that (i) perform calculations with numbers, (ii) manipulate or process data (information), and (iii) control continuous processes or discrete devices—often called manipulators—in real time or pseudo real time.

Within this rather broad definition, some of the key characteristics which epitomize the young breed of electronic, digital, stored program computers are:

1) *Their digital operation*, which has placed great emphasis in mathematics on discrete computation methods and the use of discrete structures in the solving of mathematical and logical problems. This digital feature has such overwhelming advantages (*2*) that a strong trend toward total digital has occurred in equipment, especially where functions were previously implemented by electromechanical or analog devices.

The digital feature—often considered a limitation—of the computer has necessitated the conversion to discrete form of the previously more common continuous world of curves, surfaces, and functional relationships with which mathematicians, physical and biological scientists, and economists were most familiar. As Birkhoff so aptly put it (*3*), "computers seem to have restored a balance in mathematics. Newton and Leibniz along with the Calculus emphasized and focused attention on continuous functions and theories of convergence and limits. Computers spun attention around to discrete mathematics. Combinatorics, heuristics and number theory were exemplary of the fields of mathematics that have flourished since the advent of computers."

The computer's digital nature was due in turn to the nature of electrical circuits circa the 1940's when the most common form of storage of numbers (data) were the flip-flop, the electromechanical relay, and the gas tube. These devices which could "exist" in one of the two states led to the need for binary representation of numbers. The rest is history. We have become accustomed not only to binary representations of digital numbers and alphabetic characters, but we now are quite comfortable with curves, graphics, and pictures formed by binary means, that is, through the placement of dots (or spots of light).

The many advances in electronic and optical technology have not changed our reliance on digital operation or discrete representations of relationships and computation methods. The large-scale integrated (LSI) circuits of 1976, which contain on chips less than 1-inch square, the equivalent logic found in 1956 in a room full of vacuum tube circuitry, are still dependent on being able to distinguish between one of two electronic states and thus on binary operation.

2) *Their stored program capability*, probably the most distinctive feature of the new computer of the 1950's and the single characteristic which has led to the so-called "Second Industrial Revolution," or the computer revolution of the latter half of the 20th century. This idea—an insight of genius—was that of John von Neumann. He conceived and described in the 1940's the notion that the instructions for the computer be written in the same form as the data being used by the computer; then, instructions could be operated on just as could numbers; this allowed instructions to be modified and different sets of instructions to be selected based on intermediate results derived by the computer with no human intervention. This idea as expressed in its initial conceptual stages is interesting. In 1946 we find von Neumann, Burks, and Goldstine writing (*4*, p. 1):

> . . . *the machine must be capable of storing* in some manner not only the digital information needed in a given computation . . . and also the intermediate results of the computation . . . but also *the instructions which govern the actual routine to be performed* on the numerical data. In a special-purpose machine these instructions are an integral part of the device and constitute a part of its design structure. *For an all-purpose machine it must be possible to instruct the device to carry out any whatsoever computation that can be formulated in numerical terms* [italics mine].

Coincident with the concept of the stored program was that of the computer itself being able to select and vary the sequence of instructions carried out. Again, this remarkable conception was due to von Neumann and in early writings (*4*, p. 3) was described as follows:

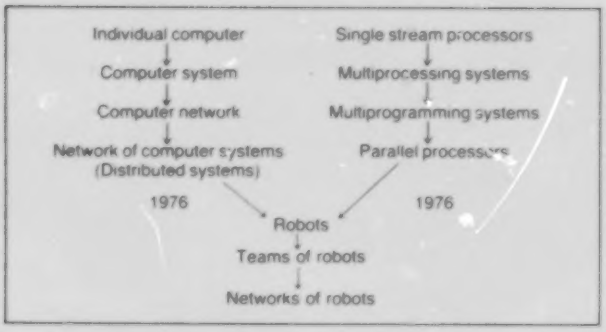
> The utility of an automatic computer lies in the possibility of using a given sequence of instructions repeatedly, the number of times it is iterated being either preassigned or dependent upon the results of the computation. When the iteration is completed, a different sequence of orders is to be followed, so we must, in most cases, give two parallel trains of orders preceded by an instruction as to which routine is to be followed. . . . Consequently, we introduce an order (the conditional transfer order) which will, . . . cause the proper one of two routines to be executed.

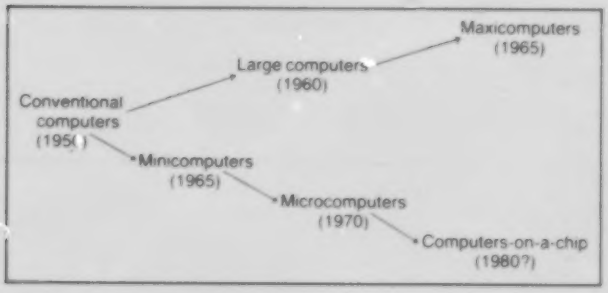
Thus, the truly innovative concept of the stored program is dependent on the realization of conditional transfers and indexing internal to the computer and free from the need for human intervention.

3) *Their self-regulatory or self-controlling capability* (*5*) which is based on repeated sequences of prediction, performance, observation and measurement of the difference between actual and desired result, and modification allowed purposeful behavior to be engineered in hardware. This capability, incorporating, as it did, the concept of feedback led to computers probably being the first "goal-seeking" machines and the first robot devices. It is apparent that the capability for self-controlling or self-regulating is dependent on automatic modifiable stored programs (see point 2 above).

4) *Their automatic operation*. Although this capability may be derivative from the computer's modifiable stored

The author is director of the Institute for Computer Sciences and Technology, National Bureau of Standards, Washington, D.C. 20234.





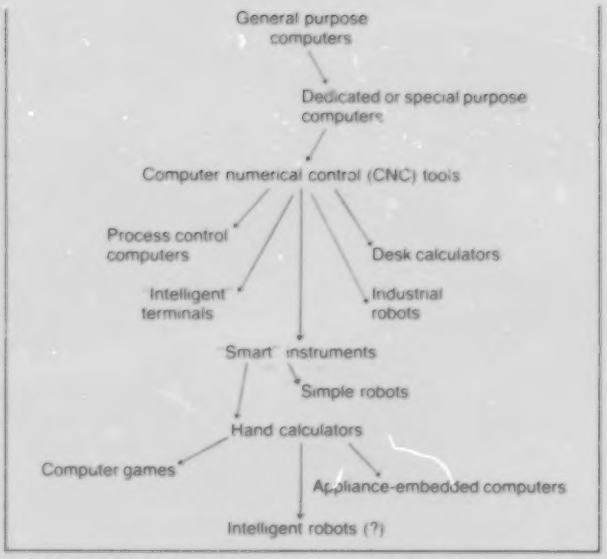


Fig. 1 (top left). Engineering developments highlighting complexity of interconnection or function. Fig. 2 (bottom left). Engineering developments highlighting the size of computers. Dates in parentheses approximate the year in which the particular computer type was first commercially available. Fig. 3 (right). Engineering developments highlighting the packaging of computers for end-use or for application (in approximate chronological order).

program and self-regulatory features, it deserves separate identification. The capability of automatic operation, which implies that the computer operates independently of human operators and human intervention after its computations start, is essential to the early visions and theories of von Neumann, Norbert Wiener, Claude Shannon, and Alan Turing. It is also the basis for the more sophisticated applications of computers in space, nuclear energy, weapons delivery systems, process control, cryptography, and robotics.

5) *Their reliance on electronics.* This is perhaps the most time-dependent feature of computers and it may be a significant trait only for the first 25 years of history. And yet the use of electronic devices (6), in which the movement of electrons could be controlled, provided for the first time in the 1940's and 1950's the reliability and speed of operation, the practical size, and the automatic operation that made the computer as we know it today possible. Electromechanical relays, hand-wired plugboards, and manually set switches, as opposed to electronic circuitry, simply would not have provided an adequate engineering basis for computer development.

By the 1970's however, electronics had been joined by optics, by crystallography, by plasma physics, and even by polymer chemistry in providing the components for computers. Indeed, the need for materials and packaging brings with it all kinds of complex problems of material processing and material choice in semiconductor devices. Also, surface physics and chemistry are becoming of extreme importance as miniaturization, plasma etching, and superconductivity become part of the computer design process (7).

In summary, then, the principal traits of computers which have been reflected in their evolution to date and which hold the key to immediate desired advances are (i) their digital operation, (ii) their stored program capability, (iii) their self-regulatory or self-controlling capability, (iv), their automatic operation, and (v) their reliance on electronics.

Environment Generated by Computer Engineering Developments

Engineering developments since 1950 have not converged to yield a specific instrument that typifies a computer. One might expect that this would have occurred on the basis of historical precedents such as the automobile, the airplane, the microscope, the television set, and the like. However, in the computer field, engineering and architectural advances coupled with almost revolutionary changes in mathematics and logic have actually resulted in a divergence in what might be considered a computer.

There are several pragmatically distinct lines of engineering development which display the rapidly changing computer environment of the last 25 years. They can be pictured in terms of (i) complexity of interconnection and function, (ii) size of computer device (or system), and (iii) packaging for end-use or for application.

Graphically, these developments are shown in Figs. 1 to 3, where each listed entry represents a physical manifestation of a computer, a group of computers, or a self-contained computer-controlled device.

In Fig. 1 (8), *single-stream processors* are confined to the sequential or consecutive execution of program instruction streams; *multiprocessing systems* are systems that will simultaneously execute two or more computer programs; *multiprogramming systems* are systems that provide for the interleaved execution of two or more computer programs by a single central computer; *parallel processors* allow for the simultaneous execution of two or more process streams in a single unit; and *robots* are defined as devices that have computerized motor control, computer-controlled sensor abilities, and some features of human intelligence, and that interact directly with their physical surroundings.

In addition, it is important to note that (9),

> A computer network, in the broad sense, is any system including both data processing equipment and data communications equipment. . . . The data processing equipment

may consist of some number of computers or a variety of different terminals, supplied and maintained by the users of the network. The communications equipment may include a mix of equipment such as circuits, modems, multiplexors, and concentrators.

Computer technology has been changing so rapidly that definitions for the computer types presented in Fig. 2 are often based on cost rather than on engineering characteristics. However, widespread use of these terms makes it worthwhile to have definitions for reference. The following are suggested (see *10*).

▶ *Minicomputers* are a "class of stored-program digital computers suitable for general-purpose applications and are priced below $50,000 in their minimum configurations." A typical minicomputer is a parallel, binary processor with a 16-bit word length, weighing less than 50 pounds, consuming less than 500 watts of standard 115-volt electric power, and requiring no special air-conditioning.

▶ A *microprocessor* is the central unit of a microcomputer. It is not the entire computer, but it contains the logical elements for manipulating data and performing arithmetic or logical operations on it. To make a complete computer, the microprocessor must be augmented through several support elements such as memory, input/output (I/O) circuits, and other specialized functions.

▶ A *microcomputer* contains at least one microprocessor plus other supporting circuitry, such as I/O interfaces, DMA (direct memory access) logic, interrupt circuitry, memory, and real-time clocks.

▶ A *computer-on-a-chip* is perhaps best described by noting that (*11*) by the early 1980's, it should be possible to visualize the existence of a complete minicomputer system chip less than an inch on a side including a 16-bit central processor unit, 32 kilobits of memory, and simple I/O interfaces. Speed, as limited by power consumption and serialized I/O, might be in the range of 10^5 to 10^6 instructions, and manufacturing costs should be $10 or less.

Most of the names in Fig. 3 have by now also become part of everyday jargon. For example, process control computers control continuous manufacturing processes, such as chemical processing, as opposed to discrete manufacturing processes; "intelligent" terminals are office—or home-size—terminals that interconnect with computer networks and contain sufficient logical circuitry to perform like a computer in a freestanding mode; simple robots are exemplified by the Viking lander robot and the lunar robots of the U.S.S.R.; computer games have appeared in the last 3 years in the commercial market; one popular game uses the home television screen as an electronic display device and contains the controls needed for playing games similar to ping-pong, handball, racquet ball, or squash. Appliance-embedded computers are recent developments using inexpensive, miniaturized logic chips to provide computer control of familiar appliances. Recent highly publicized examples are the widely advertised microwave ovens of 1976 controlled by microprocessors. Finally, intelligent robots, possessing some features of human intelligence and capable of performing some functions previously done only by humans, are close enough at hand to have incited preventive actions.

From these figures, observations and trends which bear both on the evolution of computers and of computing can be extracted. Some fall most naturally under topics appearing subsequently; a few can best be made here.

1) A cominant trend has been the decrease in size and cost of cómputers (for example, minicomputers, microcomputers, and computers-on-a-chip), which places tremendous computing power within the reach of individuals. At present, this wealth of individual computer assets is exemplified by hand calculators, ranging in price from $10 to $400, and microprocessor kits typically in the price range of $50 to $500. This availability to individuals of computer power under their own control is already exhibiting real potential for dramatic change in the providing of education; in the marketplace for engineering services; in the type and use of home appliances with potential impact on the role of homemakers; in the provision of health care and in the possibilities for really effective self-practiced preventive medicine.

2) The increasing attractiveness of computer networks has placed almost unlimited computer power in the hands of groups and organizations no matter how small, and in the hands of individuals. The initial price is that of an inexpensive terminal and the telephone lines by which it is connected to computer power "for sale" literally anywhere in the world. The sale price of this computer power is generally based on the computer time and number of computer components used and can, of course, be extremely expensive. But the means are in place through computer networks. Some of the more striking attributes of computer networks include (*12*): (i) The sharing, nationally and internationally, of expensive information resources, scientific computational resources, and computer equipment. (ii) The possibly best means of providing equality of access to and an equality of quality in public services, independent of geographical location. Primary examples are those of health care in remote areas, educational services, particularly continuing education, and public protection through law enforcement networks. (iii) The centralized management, in a real-time sense, of geographically dispersed organizations, whether they be governments, multinational corporations, worldwide weather services, or worldwide monitoring services protecting against natural or man-made catastrophes. (iv) The "perhaps" only means for resurrection of the entrepreneurial environment on which so much of our national industrial and scientific creativity and innovation depends, and which itself depends in turn on the availability to the entrepreneur of the most current information and most modern computational (research) assets.

3) The availability of maxicomputers with sophisticated parallel processing capabilities and complex interconnectivity of components has provided an unprecedented enormity of scientific, statistical, economic, and engineering calculation power. Calculations, projects, and problem solutions totally beyond human attainment which have and are occurring through the use of computers include long- and short-term weather forecasting with useful accuracy, nuclear reactor design and control, economic modeling, modeling of socioeconomic systems for policy purposes (*13*), control of weapons delivery systems, the U.S. manned space program, environmental monitoring, the direct-distance-dialing system of AT&T, and cryptographic applications.

4) The rise of the "computer hobby shop" will probably provide the strongest force to date for developing a first generation of individuals who feel comfortable with computers, who are not afraid of them, and who will exercise in the computer arena that famous American "basement creativity" that has boded so well for us as a nation in the past. The computer hobby shop is not a new phenomenon for those of us who were at least of teen age at the end of World War II. In the early to mid-1950's, military electronic surplus stores sprang up in cities like toadstools. Some stores even

sold surplus or secondhand electronic components by weight.

From this superb collection of electronics—the product of hundreds of millions of dollars of government and industry R & D—amateurs learned to build oscilloscopes, electronic repair equipment, and amazing electronic gadgets. The electronics surplus stores in the 1950's bred the generation of engineers that was able to conceive and advance the computer technologies, the TV technologies, the satellite technologies, and the space technologies of the 1950's, 1960's, and the 1970's.

The computer hobby shop of the 1970's and probably the 1980's is today's counterpart of that older electronics surplus store. In these stores, you can buy teletype machines, modems, acoustic couplers, older versions of electronic display screens, adder circuits, bits and pieces of computers, and all kinds of microprocessors and large-scale integrated chips.

The computer hobbyist of today is any age—from teen-ager to retiree. But for what new science, technology, or engineering fields this new hobbyist will be the pioneer, we do not know for sure. However, I am willing to venture a few conjectures; namely, that our near-term automation and electronics future will see computer control systems in most of our home appliances, our automobiles, and our personal work-easing contrivances such as lawn mowers and powerboats.

A fair prediction for the future based on trends in computer engineering is that there will be a heavy emphasis on computer control of continuous and discrete processes and of real and near real-time processes. This computer control "boom" will result both in substitution of computer control for traditional control systems (as is already occurring in automobiles) and in the invention of entirely new processes not previously possible without computer control. Anticipated instances here include those of useful working robots and of extraterrestrial explorations and colonization.

Quantification of the Evolution of Computers

One of the surprises in reviewing the evolution of computers is the grasp one obtains of the advances, uses, problems, and impacts of computers and computing by noting where computers are and who has been using them. For that reason, some historical and current data are displayed and assessed below.

In the early 1950's, almost all computers were owned by or devoted to applications of the federal government. For example, the first two electronic, digital, stored program computers were ordered by the government to handle the 1950 census. In rapid succession, computers were ordered by the government for weapons calculation, weapons delivery and control, cryptographic applications, nuclear reactor design, nuclear engineering, and inventory and logistical applications.

The computers of 1950 were not devices. They were rooms—big rooms—full of tubes, circuits, ventilating equipment, and people. I might add that the people carrying tapes, pushing buttons, and lowering the room temperature in these old monstrous computers were just as important a computer component as were the tubes and the circuits.

In the mid-1950's, there were fewer than 1000 computers in the United States, and they were all conventional computers with continuously increasing power. There were fewer than 100,000 computer professionals, and about 100 technical leaders in the field in the late 1950's, most of whom knew each other well. In the mid 1960's, there were about 30,000 computers—still conventional big computers. We knew where they were and what they were doing.

At the end of 1976, instead of 1000 conventional computers in the United States, we had some 220,000 computers, about 40 percent being medium or large computers and 60 percent being minicomputers. Minicomputers (minis) are small in size and by definition cost less than $50,000. Most cost less than $20,000. A lot of them cost less than $10,000. As the number of computers increased from 1000 to 220,000 in 20 years, and the number of computer professionals—analysts, designers, programmers, and operators—increased from about 100,000 to about 2,500,000, the number of actual users became impossible to count—especially as most computers became part of computer networks. An easy calculation tells us that, with 220,000 computers in the United States and about 220 million people in the United States, there is now one computer for every 1000 people or one computer for every 240 families (54 million families in 1975). That is about one computer for every good-size high-rise apartment building or every 240 houses.

And we have not yet mentioned microprocessors or microcomputers. As the number of minicomputers and conventional computers increased from 30,000 to 220,000 in the 10 years from the mid-1960's to the mid-1970's, the number of microprocessors increased from none to three-quarters of a million.

By 1980 the number of minicomputers will reach about that number (750,000) but the number of microprocessors will be more than 10 million. They are and will be so small—in the range of inches on a side—and so inexpensive—$10 to $500—for central processing units and logical units that it will be more practical to buy a number of them than to test a single one for reliability.

As minicomputers became a marketable product, we saw the customers or users change from large institutions to small institutions and wealthy individuals. With microprocessors, the customers are potentially everyone.

In addition, it is estimated (*14*) that some 30 percent of the computers in the United States are parts of computer networks (that is, they use an average of 20 terminals) and that (*15*) some 34 percent of federal computers are part of networks.

These statistics and data are useful in that they make it apparent that we can no longer know who is using computers and what they are computing. Indeed, when anyone with a telephone and between $25 to $50 per month or $2 to $5 per hour can rent a computer terminal and have access to a computer, the number of users is realistically indeterminate.

Other significant information is directed toward finding out who is using computers. As was mentioned above, the federal government was the principal owner and user of computers during the 1950's. That picture has now changed dramatically, as can be seen in Table 1, which shows the ownership within the United States in 1976 of conventional general-purpose computers (not mini- or microcomputers). We can surmise that the future evolution of computers and computing will be markedly influenced by this new mix of ownership, just as the initial phases of development were so strongly influenced by the federal government.

Programming and Software Development

It is very dangerous to venture an opinion that there is "something new under the sun." However, I would conjecture that programming, software, and

flow-charting were essentially new concepts tied entirely to the advent of computers. It actually appears possible to assign 1946 as the year the flow chart (or flow diagram, as it was called then) was "invented" by Goldstine and von Neumann (*16*). Even then, the inventors correctly perceived the flow chart as a "logically complete and precise notation for expressing a mathematical problem" which evolved as a "geometrical drawing to indicate . . . the iterative nature of an induction." As a still essential element of programming, Goldstine's statement (*16*, pp. 266–268) of 1946 which comments that the purpose of the flow chart is to give a picture of what is going on in the control process of the computer as it solves a problem, to show the states of variables at key points in the course of the computation and to indicate the formulas being evaluated, is still valid.

In the same manner, 1947 can be assigned as the date for the invention of the computer program as we know it today. Since 1947, programming has not changed in its basics. The term "software" has been coined to include all the programs needed to operate the computer, to perform a given application, and to allow individuals to communicate with the computer.

During the period from 1947 to the present, software or computer programs have evolved into a number of separable product lines differing in purpose and in market availability. Also, as might be expected, a software engineering language or jargon has evolved which has become part of the language of computer science and engineering. About 1968, the phrase "software engineering" itself was adopted (*17*) to describe activities such as programming methodology, software development, programming tools, and programming standards. It is still a young, diverse, disorderly, and expanding field. It is possible, nevertheless, to pictorially show the principal generic product lines into which software has been fitted (Fig. 4).

As one proceeds from left to right in Fig. 4, the software product lines become less dependent on or allied to particular computers and more oriented toward user needs. Programming languages, for example, have been developing primarily since the late 1950's to simulate as closely as possible the natural languages which people use. They employ English words and well-known mathematical symbols as much as possible.

Programming languages are probably the most significant development in the history of software. They allow anyone with just a minimum of training to write or to use computer programs. As a result, they have made computers a universal resource rather than just a device for scientists and engineers.

Table 1. Ownership (by percentage) of general-purpose conventional computers within the United States, 1976 (14).

Ownership by industrial classification	Percent of computers
Manufacturing industry	31.0
Electric machinery—3.5%	
Nonelectric machinery—4.5%	
Other process manufacturing—9.7%	
Other manufacturing—11%	
Transportation equipment—2.3%	
Miscellaneous business	13.3
Advertising, employment, equipment, rental, engineering services, other professional services)	
Banking, credit, insurance, real estate, and other financial institutions	
Trade (wholesale and retail)	13.1
Educational institutions (schools, universities, libraries)	5.7
State and local government	5.7
Federal government	3.4
Transportation carriers	2.9
Medical and health services	2.7
Printing and publishing	2.4
Communications	1.9
Utilities (electric, gas, and sanitary services)	1.6
Other professional services	1.9
Petrochemical industry	1.0

By 1976, it was estimated that at least 85 percent of computer users were working with programming languages. The better known are COBOL, FORTRAN, BASIC, ALGOL, PL/I, and APT.

The first 25-year history of software has not been marked with the successes, the advances, or the decreasing costs associated with computer equipment or hardware. Software or computer programs have evolved from the fields of mathematics and logic; programs are essentially algorithms stated for computer solution, and the correctness of computer programs is based on methods of proof for algorithm correctness.

Listeners always evince genuine surprise when they are told that, today in 1977, there is no theoretical (or mathematically rigorous) way to prove programs correct (except for trivial programs containing less than 100 statements). This serious limitation will be difficult to overcome because of the mathematical difficulty of constructing the necessary inductive assertions and the cost of the computer time to generate the proofs (*18*). The impracticality of exhaustive testing of all program input values can be shown by observing that it would take the "fastest" machine available today more than 30,000 years to try all inputs to a simple multiplication program.

The lack of theoretical proof of correctness of computer programs has resulted in the expenditure of considerable intellectual and physical resources in the software field to develop engineering and statistical substitutes. Software engineers and computer scientists have concentrated on quality-control techniques for software development, "debugging aids," automatic programming techniques, software validation, and the like, but to little avail. Software correctness still remains the most elusive goal of computer science.

As a result, software is the most unsafe, the least understood, and the most expensive component of total computer system costs. Software development costs are now almost 90 percent of total computer system costs. This percentage will probably increase along with the absolute costs of software, since software design, development, and testing is the most highly labor-intensive component of computer system products. The really useful and exciting advances in computing will probably only proceed at the same pace as advances in software engineering. And, this will be distressingly slow.

Computing and Applications

As was mentioned in an earlier section, the evolution of computers exhibited three distinguishable generic lines of computing or application: (i) scientific calculations, (ii) data processing or information handling, and (iii) the control of continuous processes or discrete devices—often called manipulators or effectors.

This distinction is used simply for clarity in describing applications and in highlighting the involvement of different scientific disciplines and customer groups. It is not intended as a means of delineating or defining the market. For, in the current antitrust suit of United States versus IBM, great weight has been given to defining the market and to determining whether there is indeed a useful distinction to be made in that sense between scientific computer applications and data processing applications.

For our purposes in viewing the evolu-

tion of computing, it is helpful to observe the following.

Scientific calculation applications are characterized by relatively small amounts of input and output and large amounts of data computation which is often complex in nature. Scientific users have customarily demanded high calculating speeds, high performance, and great computer power. They have not traditionally placed much of a demand on sophisticated applications software because they have the scientific ability to produce their own software (*19*), especially as they gain experience with programming languages. The demands for scientific calculation projects have provided the principal motivation for the development of large, maxi-, and supercomputers. Von Neumann's interests were primarily in scientific calculations using computers, and he can appropriately be considered its first inventive leader.

Data processing applications are characterized by large amounts of input and output and relatively small amounts of data computation. Historically (*19*), data processing users have had lesser capability to develop their own computing applications and have relied more on vendors to provide to them a total computer system including the equipment, software, and continuing support services. They are, as a result, almost completely dependent on programming languages and on the availability in the marketplace of completed application program packages.

The better-known types of data processing application packages include Data Base Management Systems (DBMS) and Management Information Systems (MIS). Other familiar application packages include account billing systems, payroll systems, and inventory control systems.

The initial thrust for data processing applications probably was due to marketplace forces rather than to user demand. At least two computer systems manufacturers, IBM and Burroughs, entered the computer field as a transition from the electric accounting machine (EAM) market. It was natural to simply transfer customers' familiar data processing applications from EAM equipment to computers with relatively little regard for any of the distinctively useful features of computers. This historical occurrence appears to have indeed set the stage for the first 25 years in the evolution of data processing applications.

Computer control applications are less familiar and probably less understood

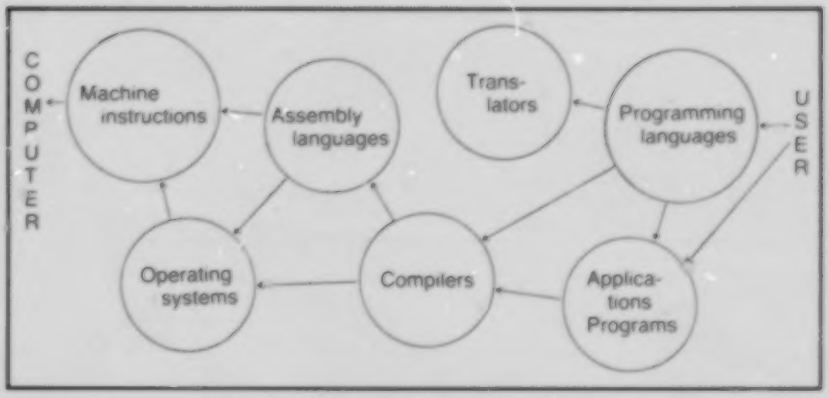


Fig. 4. Generic software product lines.

than the other two generic application product lines. In a way, this is surprising since the well-known scientist Norbert Wiener and his scientific child "cybernetics" are so closely associated with automatic control systems (*20*). Computer control applications are characterized by computer control of the state of a continuous process or of the movement and position of a device in accordance with input signals descriptive of the environment of the process or device.

Applications of this type are newer and fewer than those for scientific calculation or data processing. Examples of process control include computer controlled chemical processes, power plants, and surface mass transit systems (such as BART in San Francisco). Even more scarce are examples of devices controlled by computers. We can cite as familiar instances the Viking lander (*21*), the lunar rover, and numerical control tools.

Computer control applications highlight the importance of the self-regulatory and self-controlling capabilities of computers. Software developments are also fewer in this application area. However, there is a growing consensus that this area may be the fastest growing in the near future.

Concluding Comments

In spite of its relative youth in the history of science and technology, the computer has exerted a powerful and complex influence on our lives. In return, society and the individual, through their visions and concerns, have applied strong and widely varying pressure on computer developments and applications. Where the human has encountered the computer, there has been a play of forces and fears not yet understood or measured sufficiently to foretell the future. But a future for both computer and man, there will surely be.

A profound admiration of and concern in computers by people is to be expected. Computers are properly cited as the first as well as the most important invention ever that significantly extends man's intellectual capabilities. Until the age of computers, inventions had primarily extended our muscular powers as well as certain of our sensory powers. Extensions during the Industrial Revolution epitomize the replacement or extension of man's muscular power. Telephones and microscopes are excellent examples of extensions of our vocal and visual powers.

But throughout all history man was never threatened by anything—animate or inanimate—that could equal, extend, or surpass his intellectual capabilities, until the computer. Now in the last 25 to 30 years, there has cascaded down upon us in an electronic jumble, big computers, supercomputers, "smart" terminals, industrial robots, minicomputers, microcomputers, computers-on-a-chip, teleoperators, and real robots almost like those of science fiction.

We have seen, in this article, some of the history, the trends, the users, and the problems that have coalesced to produce the computer world of 1976. My belief based on the experience of us all is that the balance of power and the ratio of intelligence between man and computer is still indeterminate (*21*). Further, it is not entirely under man's control. In particular, as computers increase their capacities to perform more of the tasks formerly considered only within man's intellectual province, man must have and finally has, the opportunity to equip himself for other functions and a superior existence, or else his survival will seem less important to himself, leading

quite unnecessarily to a physical and intellectual ennui.

One way to contemplate the future which does not preempt our choices and which simultaneously puts the computer-man relationship in better perspective is to view the future in terms of four possible scenarios: (i) man without computer, (ii) man with computer, (iii) computers without man, and (iv) computers against man.

Within the setting of these scenarios, what we know about the evolution of computers to date leads me to predict that:

1) Progress, or changes, in the advanced, imaginative uses of computers will be despairingly slow—certainly much slower than in the first 25 years of computer development.

2) The decreasing costs and decreasing size of computers and logical devices will put these scientific artifacts into the hands of large numbers of individuals. We will see spurts of that "basement creativity" for which Americans are so renowned. Computer-related advances will be many, random, and beneficial. However, because of constraints to progress, these advances will be localized without large-scale diffusion.

3) Man coupled with computers will outlast man without computers. This surmise appears reasonable for both individuals and groups in society.

4) Man will continue to increase the number of "intelligent" tasks for computers faster than he does for himself. This paradoxical situation prevails because (i) we like to coax the utmost out of our most fascinating invention; (ii) we are intellectually lazy; and (iii) many people already feel insecure without the computer and are, in short, productively addicted to it.

5) Computers will provide to the individual more control over his personal environment than he has ever before been able to exercise. This capability will result from the miniaturization of computer components along with the decreased cost of computer hardware. Controls now possible range from individual self-paced instruction to the monitoring in real time of vital health signs, to surveillance for public protection of dangerous surroundings such as alleys and hallways. Attainment of these kinds of individual control will be accomplished principally through local ingenuity.

6) Major efforts will be directed toward the use of computers for increasing public accountability. This will take the form of more computers used for more record-keeping tasks.

7) In spite of all man-made constraints, there will be an irreversible but slow trek to realize with computers, forms of intelligent behavior that are essentially limitless, transcending man and computer taken separately.

I feel quite sanguine that people will not, as is so frequently stated, become the victim rather than the master of computers. Further, we are all most fortunate to be part of the process that will be making monumentally important decisions about man's intellectual future.

References and Notes

1. H. D. Huskey and V. R. Huskey, *IEEE Trans. Comput.* **C-25** (2), 1190 (1976).
2. G. K. Kostopoulos, *Digital Methods*, (Wiley Interscience, New York, 1975), p. 1.
3. G. Birkhoff, "Applied mathematics and its future," speech at National Bureau of Standards, 26 March 1976.
4. A. W. Burks, H. H. Goldstine, J. von Neumann, *Preliminary Discussion of the Logical Design on an Electronic Computing Instrument* (Institute for Advanced Study, Princeton, N.J., 1946), vol. 1, part 1.
5. C. Eames and R. Eames, *A Computer Perspective* (Harvard Univ. Press, Cambridge, Mass., 1900), pp. 6 and 128–131.
6. D. A. Bell, *Fundamentals of Electronic Devices* (Reston Publishing, Reston, Va., 1975), pp. 1–7.
7. L. M. Branscomb, *Phys. Today* **29** (1), 55 (1976).
8. *American National Dictionary for Information Processing* (American National Standards Institute, New York, in press).
9. I. W. Cotton, "Executive Guide to Computer Networking," National Bureau of Standards, Washington, D.C., publication in process (available in draft).
10. Information supplied by DATAPRO Research Corporation, Delran, N.J., 1976.
11. D. A. Hodges, *Comput. Des.* **15** (2), 77 (1976).
12. R. M. Davis, *Computer* **6** (4), 14 (1973).
13. M. Greenberger, M. A. Crenson, B. L. Crissey, *Models in the Policy Process* (Russell Sage Foundation New York, 1976), pp. 1–42.
14. Data supplied by the International Data Corporation, Waltham, Mass., 1976.
15. General Services Administration, Inventory of Automatic Data Processing Equipment in the U.S. Government, 30 September 1976.
16. See H. H. Goldstine, *The Computer from Pascal to von Neumann* (Princeton Univ. Press, Princeton, N.J., 1972), pp. 266–270.
17. R. T. Yeh, *Computer* **8** (5), 15 (1975).
18. R. London, *Proc. Assoc. Comput. Mach. Annu. Conf.* **1**, 39 (1972).
19. United States versus IBM, plaintiff's pretrial brief 1 of 17 October 1974, Southern District Court, New York.
20. N. Wiener, *Cybernetics* (MIT Press, Cambridge, Mass., 1948), pp. 13–15.
21. C. R. Spitzer, *Spectrum* **13** (10), 92 (1976).

Large-Scale Integration: What Is Yet to Come?

Robert N. Noyce

A single large-scale integrated circuit in 1977 can contain more active elements than the most complex electronic equipment ever built 25 years ago. Enormous advances have been made in the last two decades, and we may ask whether this technology is reaching a limit or whether it will continue to advance at a breathtaking pace. By their nature, integrated circuits are small and are limited in power dissipation. As a result, their primary use is in information manipulation, not in areas where high power is required, such as in transmission or in servo drivers. However, a scanning of the titles of this issue shows that information-handling or computerlike devices are central to electronics today. Thus, large-scale integration can have a major influence on the direction of electronics in the future.

The most obvious effect of the development of integrated circuits has been to reduce the size of electronic equipment. However, the principal advantages that have accrued lie in the reduction of cost and in the improvements in performance, reproducibility, maintainability, and reliability.

Interconnections Are the Problem

Interconnections have been, and are, the major problem in computer hardware. Because of the physical size of these interconnections, delays are incurred in transmitting signals from origin to destination with the computer. Because of the capacitance of the interconnections, a power loss is incurred every time a signal is imposed on the

The author is chairman of the Intel Corporation, Santa Clara, California 95051. This article was adapted from a forthcoming book on the future impact of computers and information processing, M. L. Dertouzos and J. Moses, Eds., MIT Press.

line, necessitating high-power drivers in the system which are not needed to drive the actual logic of the computer. And because the interconnections themselves act as antennas, they transmit to and pick up from other interconnections unwanted signals, or noise. As a result, there is an exploding design of increasing power, increasing noise, and increasing size, which requires higher power, longer delays, and higher cost. With current technology, the essential logic and memory of a large modern computer could be contained within 10 to 100 cubic centimeters of silicon chips. Interconnections add not only to the size but to the apparent complexity of the large computer, resulting in maintenance, documentation, and reliability problems that increase in severity with size. The integrated circuit is the component industry's solution to the interconnection problems.

If we consider only the cost of the active elements in the integrated circuit, we quickly conclude that the single transistor is the most cost-effective.

A simple model would say that if a particular slice of silicon yielded chips of which only 10 percent were good because of the inclusion of random defects, then if it were used to make chips of twice the area it would yield only 1 percent good chips. (Actually the situation is not this bad, since defects are not randomly distributed.) The cost for twice the function then would be 20 times as great (twice the processing cost since twice the area of silicon is used, and ten times the cost because of loss of yield). Clearly, if the cost of the active silicon were dominant, such a doubling of complexity would not be cost-effective; carried to an extreme, the single transistor would be the most cost-effective.

The other major cost borne by the component manufacturer is that of assembling and testing the devices. Assembly is the process of putting the tiny silicon active element in a housing, which includes a mechanical transition from the microscopic interconnections of the integrated circuit to the sizes normally encountered in electronic equipment—that is, from a lead separation of 10 micrometers to one of 2500 μm or 2.5 millimeters. As a first approximation, assembly costs are independent of the function included on the integrated circuit chip, although they will increase somewhat as the number of electrical connections to large chips increases. Similarly, test costs increase much more slowly than the complexity of the chip being tested, although very sophisticated test equipment is required to achieve this result. Thus, the total cost per function will be made up of two elements: one increasing with complexity, of the general form ae^{bN}, representing the cost of the silicon chip; and the other, of the form c/N, representing the cost of assembly and test, where N is the number of functions included. This cost will have a minimum, as indicated in Fig. 1. As processes for manufacturing integrated circuits have been perfected and yields of good circuits have been improved, this minimum cost point has moved to circuits of higher complexity. It has been the strategy of the semiconductor device manufacturers to supply circuits that are near this minimum at any given time.

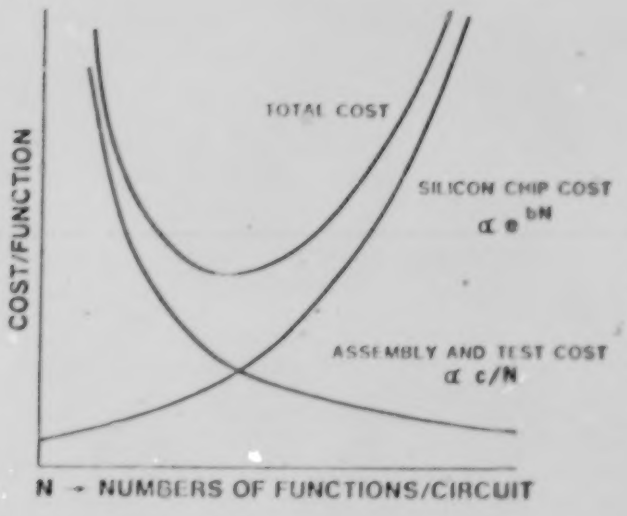
An examination of the integrated circuits offered by the industry as a function of time provides an estimate of how the complexity of products at the minimum cost point has increased with time, even though such products are introduced before they represent the lowest cost per function point. As indicated in Fig. 2, the complexity of products at the minimum cost point has doubled every year since the introduction of the integrated circuit. If the present rate of increase of complexity were to continue, integrated circuits with 10^6 elements would be available in 20 years.

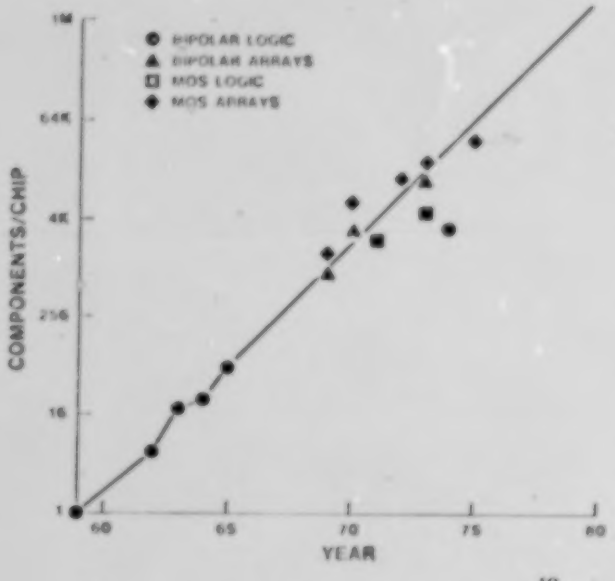
Higher levels of integration have been achieved in three ways: by increasing practical chip size through the reduction of random defects, by introducing circuit innovations that have allowed higher functional density, and by making the individual circuit elements smaller. The last has been the most important factor. We can expect the achievement of smaller dimensions to be the principal contributor to increasing levels of integration in the future. As device dimensions are made smaller, all of the device parameters move in a favorable direction, as can be seen by scaling the device dimensions and noting the changes in these parameters.

The electric fields cannot be increased as the dimensions are changed, since devices are designed with limiting electric fields now, and higher fields will cause avalanche breakdown or quantum mechanical tunneling. Thus, if x is the characteristic dimension of the device, the operating voltages of the device will vary as

$$V \propto x$$

Since the fields terminate in sheets of charge with finite thickness, the charge densities must increase to scale this





thickness. Thus the impurity densities in the device vary inversely with x. Similarly, the mobile charge n must increase as $1/x$. The area A of the conducting path varies with x^2, so the current varies as

$$I \propto nA \propto x$$

since the field is constant. The characteristic impedance of the circuit is constant

$$Z \propto V/I = \text{constant}$$

We may look at circuit delay times from two points of view, in terms of carrier transit time, τ_t, and capacitance charging time, τ_c. The transit time τ_t will be proportional to the distance traveled

$$\tau_t \propto x$$

since the field is constant. Charging the collector or drain capacitance depends on the voltage, current, and capacitance. The capacitance, C, varies as the dimensions, so

$$\tau_c = \frac{CV}{I} \propto x$$

which is similar to the variation in transit time.

The power per circuit will be proportional to the product of current and voltages

$$P \propto VI \propto x^2$$

The power density, P/x^2, is then constant, since the area of the circuit varies as x^2.

Thus, as dimensions are made smaller, all of these variables move in a favorable, or at least not unfavorable, direction. Speed increases, density increases, and power density does not increase.

Since the operating voltages decrease as the dimensions decrease, the allowable voltage drop along interconnections in the integrated circuit will decrease in proportion. A scaling argument, however, shows that these voltage drops will remain constant. For the normal metal interconnections, today's designs are far from the limits at which problems would arise. Some interconnection schemes, such as the use of polycrystalline silicon, are near the limit now. However, the use of other techniques does not appear to pose a great problem.

Figure 3 shows the historical trend in minimum average dimensions in advanced integrated circuits. Reduction by a factor of 2 is being achieved every 5 years. Were this to continue, circuit densities would increase by a factor of 64 and speeds would increase by a factor of 8 in the next 15 years.

Defect Density

The production of integrated circuits has always been a "yield" problem. Many circuits are made, the defective ones are thrown away, and the good ones are sold. Defects can arise from many sources. The photomasks used may have pinholes in dark areas, or opaque specks in areas that should be clear. Severe defects in the basic silicon crystal can make the circuit inoperative. Dust in the photoprinting operation or other processing steps that affects a critical spot in the circuit will cause failures. Misalignment in successive photoengraving steps or lack of control of critical dimensions and impurity concentrations will make the circuit inoperative.

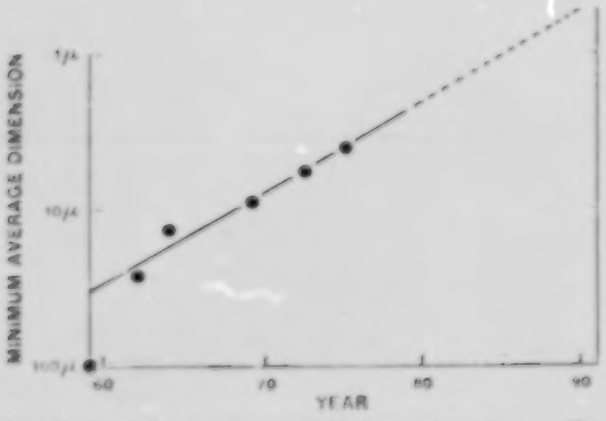
The progress in reducing defect densities can be seen from the history of silicon chip sizes of new products, as shown in Fig. 4. The absolute defect density is approximately the inverse of the chip size, since at higher defect densities the chip cost would be prohibitively high, and at lower defect densities higher levels of integration give a cost advantage.

In 15 years the extrapolated reduction in dimensions will reduce internal gate delays to the order of 1.0 nanosecond. Propagation delays will then begin to become significant if system dimensions are much more than 10 cm. During the same period, the silicon chip dimensions can be expected to increase to about 2.5 cm on a side. Thus, delays due to on-chip propagation should not be a problem, but delays in signal propagation between chips will be a significant factor.

The combined density and chip size extrapolations would result in an increase of functional complexity by a factor of over 2000 during the next 15 years. Costs would increase only slowly from those of today's complex integrated circuits, resulting in a cost reduction of 100 to 1 or even 1000 to 1. Clearly there is a motivation to try to achieve this result.

Practical Limits

Since most production technologies use photolithography to define the circuits, the wavelength of light looms large as a limit to the dimensions that can be achieved. The resolution of the printed pattern is involved, but a more difficult problem is the reindexing of successive photolithographic steps necessary to build up the integrated circuit structure. Today's techniques require that the previously printed pattern be examined (either by human eye or by automatic equipment) and aligned with the new pattern to be printed. This examination is now made with visible light, with the resolution which that implies. Realignment more accurate than about three wavelengths of light is not yet practical. The



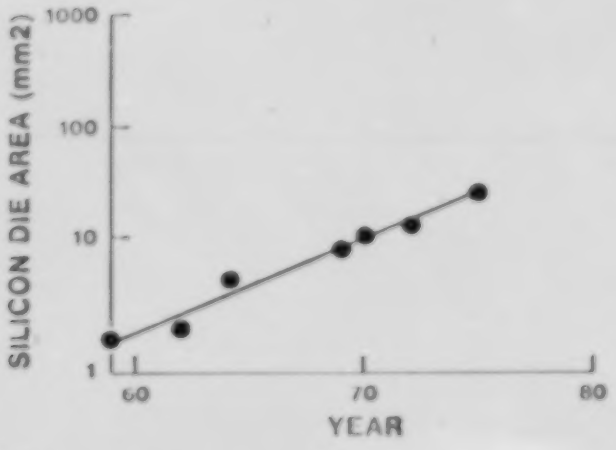


Fig. 3 (left). Historical trend in minimum average dimensions. Fig. 4 (right). Trend in silicon chip sizes in new products.

24

allowance for this amount of misalignment in the design of the circuit means that much smaller circuits cannot be produced by this technique.

One approach to avoiding this limitation is to remove the need for accurate alignment by using "self-aligned" techniques. For example, in the silicon gate technique used in metal oxide semiconductor (MOS) circuits the gate itself and the source-drain spacing are defined in the same photolithographic step. The use of self-aligned techniques eliminates the allowance for misalignment of some masking steps from the design of the circuits and allows tighter packing densities.

Another approach is to develop new device forms in which the critical dimensions are defined by techniques other than photoengraving. An example is the diffused MOS transistor, in which the source-drain spacing is defined by solid-state diffusion rather than by photolithography.

A third approach to overcoming resolution limitations is to use electron beams rather than optical techniques. Indeed, the use of electron beams for the preparation of high-resolution photomasks is now being introduced (*1*). However, production use of electron beam lithography in printing on the silicon wafers appears to be 5 to 10 years away. The many problems of inspection and realignment techniques must be solved. In addition, production-worthy equipment must be developed. The economic return on the use of the electron beam is not clear, particularly for the pioneer who assumes most of the cost of developing the technology. Therefore, the most likely scenario is that the electron beam will be used to prepare photomasks, but optical printing techniques will be used on the silicon.

There appears to be no fundamental limitation to the reduction of defect densities. The integrated circuit industry has proved very resourceful in solving these problems. Progress has been steady and seems to be determined by the amount of effort that is expended to perfect the production processes. Since the economic return is very high, we can expect the effort to continue, with results similar to those experienced in the past.

Innovation in circuit or device forms to improve densities seems to be reaching a limit, perhaps as a result of our inability to anticipate future innovations in new circuit forms. Possibilities for improvement may come from using three-dimensional forms rather than the present two-dimensional technology.

Redundancy in some form would allow much higher levels of integration than would be possible otherwise. It appears that as circuits become more complex, the overhead expense for achieving fault-tolerant circuits is a smaller proportion of the total. For example, error correction for an 8-bit word requires an additional 4 bits, whereas correction for a 64-bit word requires only 7 additional bits. A similar case can probably be made for logic. The relative cost of including redundancy is less in more complex systems. Today error correction is used in large memory systems to reduce failure rates. Partially good memory components are used in some systems by disabling faulty addresses. As new schemes are developed for the use of faulty circuits, the permissible chip sizes could expand greatly, allowing significant cost reduction with higher levels of integration.

What Is Still to Come?

The drive to higher levels of integration has great momentum and will continue. Minimum dimensions will continue to decrease, but at a decreasing rate as optical limits are approached, reaching 2 μm in 5 years and 1 μm in 15 years (Fig. 5). Scaling arguments show that speed should then increase by a factor of 4 by 1991. If defect densities are reduced as they have been in the past, chip sizes will increase by a factor of 3 in 5 years and by a factor of 25 in 15 years (Fig. 6).

As a result of these factors, components providing 65 to 131 kilobits of memory with an access time of 100 nsec should appear in 1981 and the megabit memory chip (2^{20} bits) should appear 10 years later. The use of redundancy could accelerate these times. Component costs should be comparable to those of today's memory components, or 10^{-3} cent per bit.

For nonlterated circuits such as control logic circuits, the level of integration will be lower because of the inefficiencies inherent in packing random logic. However, the level that could be achieved in 5 years would be approximately 25,000 gates, and in 10 years about 250,000 gates. These numbers exceed the gate counts of today's medium and large processors, respectively. Internal gate delays of these systems would be comparable to those of today's high-speed computers. If the amortization of design costs could be neglected, such logic arrays could be produced for less than $100 at a cost of less than 0.4 cent per gate in 5 years and less than 0.04 cent per gate in 15 years (Fig. 7).

It is still questionable whether such high levels of integration of logic circuits will be attempted. The industry has already achieved the complexity needed for today's mass markets—for example, in the digital watch or the calculator—and the profit potential that motivates these developments can only be achieved by production in high volume.

The basic technology will be developed, however, to supply the demand for

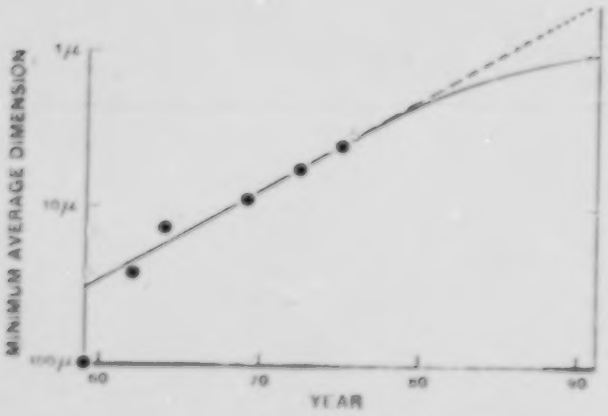


Fig. 5 (left). Projected minimum average dimension in advanced integrated circuits.

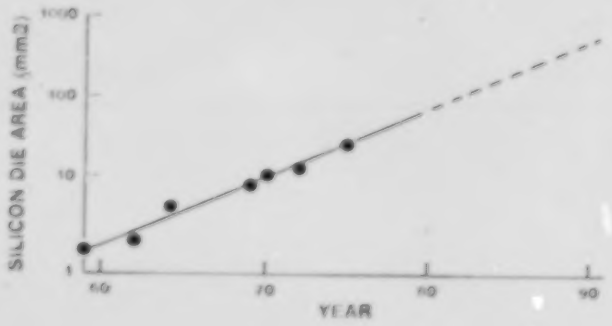


Fig. 6 (right). Projected trend in silicon chip sizes.

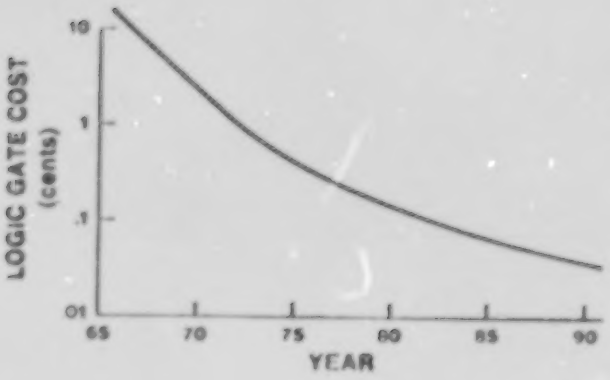


Fig. 7. Cost of logic arrays versus time.

memory. Fortunately perhaps, there is an insatiable demand for memory, and the market is completely elastic at today's usage levels. The question, then, becomes one of the cost of design of high-complexity logic for a particular application. At high levels of integration, many more unique circuit designs are necessary than for Boolean or simple logic functions. The market for each unique circuit becomes smaller, while the design costs become much higher.

The integrated circuit industry faced this problem with the introduction of large-scale integration and found a solution in the concept of the microprocessor. The microprocessor was conceived as a method of supplying many diverse requirements of random logic, thus allowing one or a small number of circuit designs to achieve a high enough market to justify their development. The true impact of the microprocessor is just being seen today, but even now it is having a radical influence on our thinking about future computing systems. We may expect the capability of the microprocessor to increase dramatically in the future, if for no other reason than that it can be done. The impact of the microprocessor is first being felt in smaller computing systems, such as terminals, controllers, and video games. As its capabilities increase, the equivalents of today's processors will be built using this concept.

Several changes will probably occur in the computer industry as a result of the microprocessor. Standardization about a small number of microprocessors will be forced by economics, since manufacturing them in volume will reduce costs substantially. It now appears that identical microprocessors will generally be available from many sources. This will not only assure competitive prices but also make possible the development of software sources independent of the hardware manufacturers. The microprocessor will become another building block, like a flip-flop or an adder, for the construction of computing systems. This suggests that architecture will change to take advantage of these very cost-effective building blocks. Increasing throughput by using identical blocks in parallel should be favored, and networks of semi-independent processors should be more cost-effective than large time-shared machines.

Structural changes will occur as well, since the microcomputer designer must consider the economics of component production and the economics of system use simultaneously. This suggests that there may be some amalgamation of the components and system producers, or at least some mutual penetration of these industries.

The potential for developing inexpensive processing power is truly awesome. With low-cost processing available, many new tasks may be undertaken which are totally uneconomical today. In addition, with more powerful processing the requirements for input-output devices may be reduced to simpler forms of information transfer such as audio input and response. Our major challenge is likely to be defining the tasks to be accomplished with this technology, and developing the software with which to accomplish those tasks.

Summary

The rapid development of large-scale integrated circuits in the last two decades has revolutionized information handling. Higher levels of integration have been achieved principally by making the individual circuit elements smaller, but reduction of random defects and innovations in circuit design have also been important factors. Practical limits to the size of integrated circuits, such as those imposed by the use of photolithography to define the circuits, can be avoided by using other methods, such as solid-state diffusion or electron beams.

The prospects are that dimensions and defect densities will continue to be reduced while speed increases. Redundancy in circuits will lead to lower costs and higher levels of integration. As logic becomes more complex and applications narrower, the increased number of unique circuit designs needed tends to increase costs. The microprocessor was developed to answer this problem and will have a radical influence on future computing systems.

References

1. R. W. Keyes [*Science* **195**, 1230 (1977)] discusses the practical limits to resolution with electron beams.

Intellectual and Economic Fuel for the Electronics Revolution

John G. Linvill and C. Lester Hogan

The past three decades have witnessed revolutions in semiconductor electronics and in electronic computers. Progress in each has stimulated and directed the other. Today, the American semiconductor industry employs approximately 120,000 people, and the foreign semiconductor industry provides employment for another 150,000. But the true importance of semiconductor devices is not revealed by such numbers. Without these devices, the computer industry would not exist in its present form and size. Many electrical and electronic devices that we accept as commonplace today—such as the hand-held calculator and shirt-pocket radio—would probably not exist at all. In this article, we focus attention on semiconductor electronics, reviewing its evolution from a unique scientific base, showing how it served and was strengthened by space and defense missions, and pointing out the remarkable development of new enterprises and new teams that formed as the revolution spread. The point of looking back, of course, is to gain insight for moving ahead. We shall look ahead by expressing predictions and imperatives for the future based on the progress of the electronics revolution to this point.

Semiconductor Electronics: Its Beginnings and Development

At the end of World War II, there was significant interest in semiconductor rectifiers and in microwave diodes used as detectors and mixers. The dominant base for the imminent electronics revolution brought by semiconductors consisted of new developments in solid-state physics and technology.

Dr. Linvill is chairman of and professor in the Department of Electrical Engineering, Stanford University, Stanford, California 94305. Dr. Hogan is vice chairman of the Board of Fairchild Camera and Instrument Corporation, 464 Ellis Street, Mountain View, California 94040.

Solid-State Physics—An Open Scientific Base

The first paragraphs of Ralph Bown's foreword to Shockley's book (*1*) give an eloquent account of the scientific base for semiconductor technology:

> If there be any lingering doubts as to the wisdom of doing deeply fundamental research in an industrial laboratory, this book should dissipate them. Dr. Shockley's purpose has been to set down an account of the current understanding of semiconductors, an understanding which incidentally is comprised in no mean degree of his own personal contributions. But he has done more than this. He has furnished us with a documented object lesson. For in its scope and detail this work is obviously a product of the power and resourcefulness of the collaborative industrial group of talented physicists, chemists, metallurgists and engineers with whom he is associated. And it is an almost trite example of how research directed at basic understanding of materials and their behavior, 'pure' research if you will, sooner or later brings to the view of inventive minds engaged therein opportunities for producing valuable practical devices. The program of work which Dr. Shockley leads was aimed at understanding a kind of materials, the semiconductors, which had already received considerable application in the communications business in the form of rectifiers, regulators and modulators. Not only were improvements in such devices hoped for but the possibility of creating an amplifier was envisioned. In the course of three years of intensive effort the amplifier has been realized by the invention of the device named the transistor.

A set of fundamental developments came from the Bell Laboratories. Single-crystal germanium and silicon were first grown there; zone refinement to provide new orders of material purity was developed. Other firsts include the junction transistor, silicon-aluminum diodes, thermal-compression bonding of metals to semiconductors, impurity diffusion, and epitaxial growth of new layers on a seed crystal.

Open communication of the new science of semiconductors stimulated intense activity. The Bell Laboratories held symposiums to inform the scientific and technical community about these developments. Publications promptly disclosed new developments.

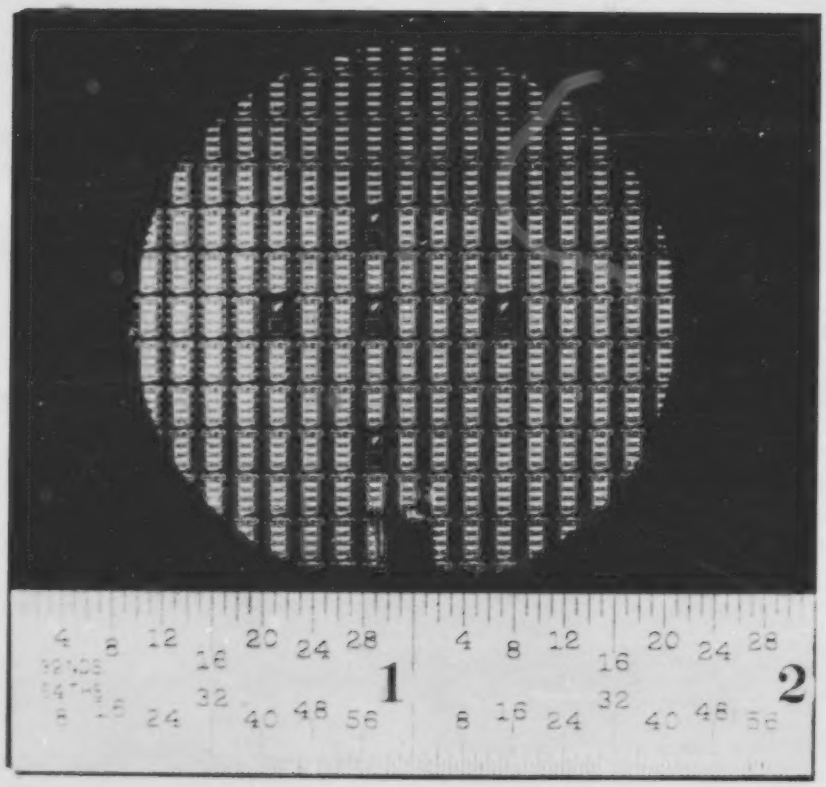
Inevitably, the gathering momentum of the semiconductor industry brought unparalleled career opportunities outside and new activities were nucleated in Dallas, Phoenix, and Mountain View. The authors, both alumni of Bell Laboratories, acknowledge the enormous and public-spirited contribution of the Bell Laboratories to semiconductor electronics in our American free-enterprise system.

Government Support of Semiconductor Electronics for Defense and Space

The importance of the transistor to defense electronics was immediately obvious. Army, Navy, and Air Force agencies immediately began the support of transistor electronics for the defense need. The pursuit of excellence was intense; competition developed among the various defense agencies to support the best ideas and the best teams in the various industrial laboratories and in the universities. Moreover, the commonality of interest among the contractors to the federal government promoted the high diffusion rate of new information in semiconductor electronics. The annual Device Research Conferences and the Philadelphia Solid-State Circuits Conferences were two forums for communication among the participants in the emerging field.

Governmental support of semiconductor R & D was large. From 1958 through 1974, various branches of the U.S. government pumped $930 million into research and development in the semiconductor industry. A very legitimate debate exists today as to whether government support of R & D in this industry has been as pivotal as the demand by military and space efforts for the devices manufactured by the industry. Very little objective data exist that can be used to support either side of this debate. However, it is our opinion that government support of university-operated research in this area was necessary in providing the atmosphere for training the brilliant students who then quickly found their places in the industry.

It is probable, however, that government-sponsored R & D in the industry itself was not as productive as the R & D which the industry chose to support itself. The reason for this is simply motivation. Government-sponsored R & D usually demands production of final devices



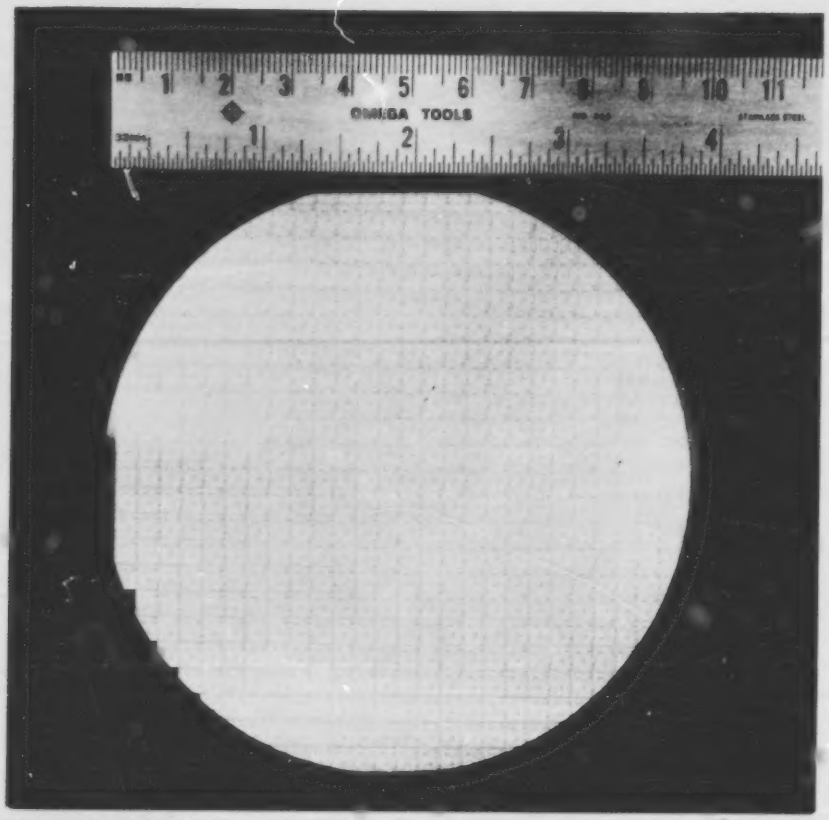


Fig. 1.(Top) Wafer manufactured in 1966; 1½ inches in diameter; 9000-square-mil dice; and 12-μm lines. (Bottom) Digital watch circuit; 4-inch wafer manufactured in 1976; 24,000 square-mil dice; 5-μm lines.

that might or might not be considered useful (that is, profitable) for the company involved. If the company chooses to sponsor R & D itself, then it is highly motivated to direct this R & D effort into products that it feels will be profitable to the company. In spite of this difference in motivation resulting in a difference in productivity, this $930 million provided useful effort that undoubtedly advanced the state of the art during its expenditure.

The U.S. Department of Commerce has published figures indicating that private industry supported $1.2 billion of R & D during the same period of time. So, all together, American industry was able to put more than $2 billion into R & D during this 16-year period. The total technical effort—including applications engineering, technical marketing, process control, and production areas—probably raises the investment that brought this industry into being between 1955 and 1975 to at least $3 billion.

In addition to R & D expenditures by the U.S. government, another important component of governmental support was contracts for production preparedness. Very early in the history of the transistor, the U.S. Army, Air Force, and Navy supplied capital dollars to the American semiconductor industry to build production equipment in order to have a capability to produce these new devices. The first such dollars were supplied in 1952 so that our industry could build production lines to manufacture alloy transistors.

While many companies were willing to support R & D efforts in this exciting new technology in 1952, few would have had the courage to build multimillion-dollar production lines at such an early date. This expenditure would have required the approval of the board of directors of the companies involved, and even if they had the foresight at that early time to begin production, their approval would have introduced a delay of at least several months. With the funds available from the U.S. government, an enterprising manager could build a production line without the approval and perhaps even without the knowledge of his board of directors.

As early as 1952 many American companies were motivated to build production lines for alloy transistors. Those who did found the business profitable. They got an early jump on their competitors who were still justifying this daring move. This technology diffused as rapidly through the United States as did R & D knowledge. Even though this ini-

tial investment by the U.S. government to build alloy transistor lines amounted to only $11 million, it provided a critical time advantage.

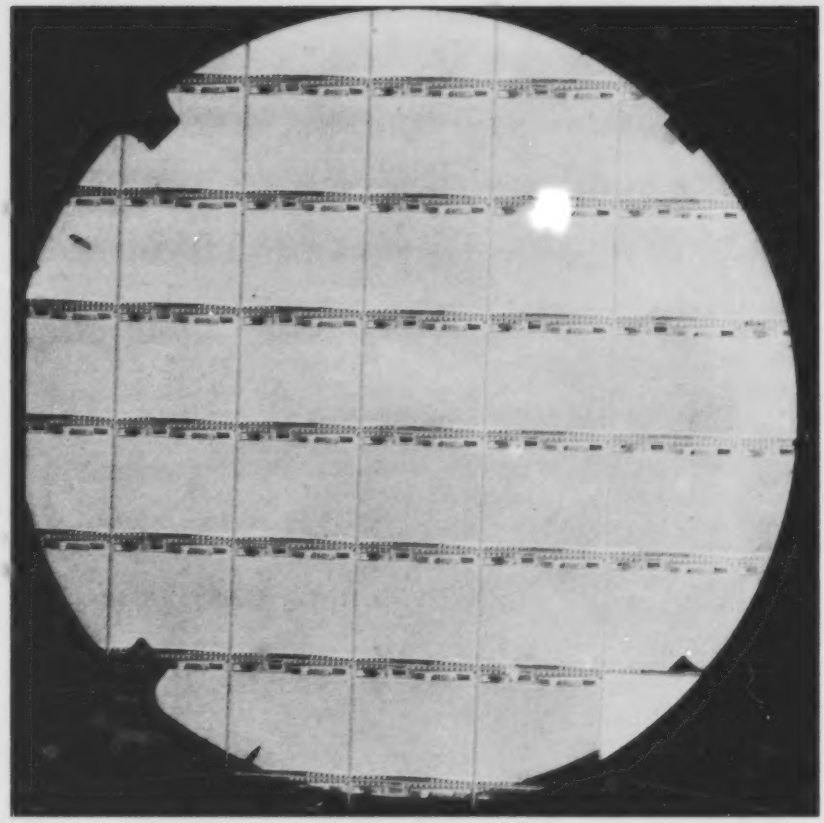
In 1957, the U.S. government provided another $15 million to industry to buy capital equipment to build production lines for diffused-base transistors. Again, while this amount was small compared to the money that private industry had to put up, it nevertheless gave an initial stimulus. As early as 1959, the services supplied another $10 million to American industry to buy or build the equipment necessary to build a production capability for integrated circuits.

Now, it is true that the $36 million supplied by the U.S. government in these contracts is a fraction of what our industry had to supply itself in order to build the enormous production capability that we now have. Nonetheless, the U.S. Signal Corps and Air Force, in particular, deserve credit for the foresight that they had so early in the game which led them to supply these needed capital dollars.

Many Americans argue that a large majority of the companies that got this early production support do not now exist as semiconductor suppliers and, hence, that the money was wasted. This is nonsense. Two of the largest recipients of the U.S. government support in the early days were Texas Instruments and Motorola. They did survive, and this early help was critical in helping them to an early start in the business. As for the companies that received support and failed, other successful companies hired their people and American industry was automatically farther along on the production learning curve because of the investment.

The largest markets in semiconductor electronics from the invention of the transistor through the middle 1960's were the computer industry and the military-aerospace industry. Two customers, because of their use volume, were absolutely pivotal in the early establishment of American semiconductor firms.

The first was IBM. In 1960, IBM was probably the largest single customer of every American semiconductor company. The second major customer was the Minuteman missile system. This missile system poured hundreds of millions of dollars into the semiconductor industry at a very important time in its history. This money went into diffused transistors and integrated circuits; in particular, it provided funds necessary for the refinements to achieve a high level of reliability for semiconductor devices. All



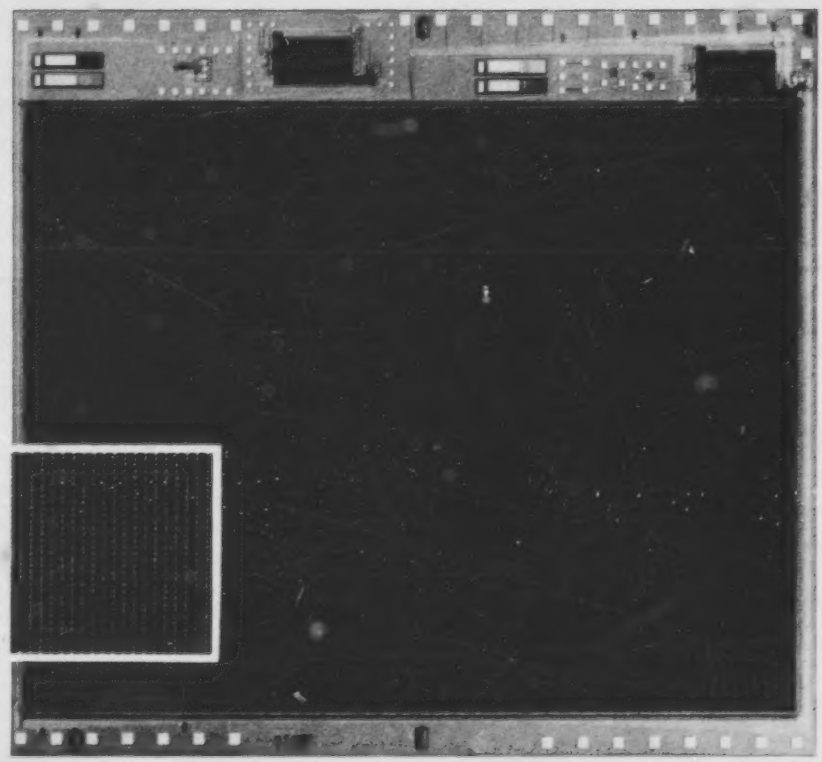


Fig. 2 (top). CCD imaging chip manufactured in 1976; 3-inch wafer. Fig. 3 (bottom). Individual CCD imaging die showing 4-μm structure.

subsequent semiconductor systems benefited from the technological advances with the new levels of electronic reliability.

Emergence of New Ventures

As electronics moved from the vacuum era to the semiconductor era, many observers expected that the industrial participants who dominated the vacuum tube component business would be the ones to establish leadership as semiconductor manufacturers. In 1955, the top ten American manufacturers of vacuum tubes in order of the worth of their sales were

RCA
Sylvania
General Electric
Raytheon
Westinghouse
Amperex
National Video
Rauland
Eimac
Lansdale Tube

The top ten American semiconductor manufacturers by sales volume in 1976 were

Texas Instruments
Motorola
Fairchild
National Semiconductor
RCA
Intel
ITT
Signetics
General Instruments
General Electric

Only two names appear on both lists. Clearly, the electronics industry in the United States exists within an economic and cultural system which not only spawned but actually encouraged the new order that was necessary for the proper winning of the vigorous new semiconductor market. The two key factors in this environment are the availability of venture capital and the mobility of American professional people. Vigorous young companies, with the motivation of partial ownership to their managers, have dominated the semiconductor industry. The American industry has been so successful that today it supplies approximately 75 percent of the world's integrated circuits.

Four separate market strategies have been successful in the development of new business connected with the semiconductor industry. The first strategy is market capture by price advantage in a widely used product. The leaders of the 1976 list (above) have all pursued this strategy with great success. They have exhibited remarkable capabilities to incorporate technological advances into their production to cut costs. They have responded rapidly to developing markets. Their performance has provided tremendous incentive to place electronic components throughout American industry in places where they had earlier been assessed as too expensive.

Some companies are of sufficient size to have sufficient internal markets for their products. Notable examples are the Bell system and IBM. Generally, the semiconductor devices produced for these internal, special markets contain similarities in some lines to the devices produced for commercial use by the large suppliers, but many devices are custom designed for the telephone and computer industries.

A third successful strategy is that of the instrument or system manufacturer which establishes and maintains semiconductor development and manufacturing capability for its internal use. The key idea here is to develop proprietary components that give the manufactured system an essential advantage over competitive systems. An outstanding case in point is Hewlett-Packard, which manufactures custom integrated circuits for its instruments, computers, and calculators. Intel, in its development of the microprocessor, exhibited a somewhat similar strategy. However, the microprocessor at this point is becoming a volume component of great importance.

The semiconductor industry, particularly on the San Francisco Peninsula, has generated a wide variety of small, successful companies which manufacture production equipment, materials, and components and provide services for the industry. Diffusion furnaces, ion implantation machines, epitaxial growth systems, mask-making systems, and others are all products of industries that are ancillary to semiconductor device production.

The Projection Forward

With the view of the dominant components of the development of the semiconductor industry to this point, we are in a better position to anticipate the developments ahead and to draw certain recommendations from them.

Centers of Action Ahead

A principal sector of the semiconductor industry is devoted to making potent microprocessors, memory chips, and standard linear circuits which can be used in enormous numbers. The last decade has seen a dramatic move from small-scale integration to large-scale integration. This change is illustrated in Fig. 1 (top) which shows a photograph of a wafer made in 1966–67 with individual dice 9000 square mils in area and having 12-micrometer lines on a wafer of 1½ inches. Below it is a photograph of a 4-inch wafer containing 5-μm lines on 24,300-square-mil dice of a 1976 digital watch circuit. Figures 2 and 3 are photographs of a wafer and die for a charge controlled device (CCD) imaging chip. The wafer is 3 inches in diameter, with each die having an area of 200,000 square mils and containing 180,000 imaging elements and 180,000 shifting elements.

There has been a great increase in the market share of integrated circuits. One decade ago, 86 percent of the sales of semiconductor manufacturers were discrete devices and only 14 percent consisted of integrated circuits. These circuits were very simple by today's standards. The most complex circuits of that time were dual J-K flip-flops with less than 25 transistors interconnected on one chip. In 1976, the share of total semiconductor sales attributed to integrated circuits had risen to 58 percent. Today, the state of the art is 16-bit microprocessors and 16,000-bit memories with approximately 1000 times the complexity of the devices built in 1966. The price for the 16,000-bit memory chip today is approximately the same as the price of a J-K flip-flop in 1966.

Both the staggering increase in complexity and the drastic reduction in cost have had a dramatic impact on the nature of the semiconductor business. In order to keep up with the rapidly decreasing prices, the industry has been forced to go from 1-inch wafers to 2-inch, then 3-inch, finally arriving today at the industry's standard 4-inch wafer. The diffusion furnaces have become larger, more accurate in temperature and gas-flow control, and much more complex.

In 1966, diffusion furnaces cost approximately $3000 per tube. Today, they cost $12,000 per tube. Today, a single mask aligner costs $50,000, an evaporator costs $80,000, 4-inch crystal growers cost $210,000, an epitaxial reactor costs $185,000. Each of these numbers is between a factor of 4 and a factor of 20 more expensive than typical equipment costs one decade ago. Ion implant equipment costs range between $100,000 and $300,000, depending on the flexibility required by the manufacturer.

Very new electron beam mask-making equipment requires a capital expenditure

of $1.5 million for one electron beam machine. A typical wafer fabrication production area in 1976 uses four times as much deionized water per square foot of factory space as it did in 1966; it uses four times as much electric power.

Some of the increase in costs of equipment between 1966 and today has been due to inflation, but most of the increase has been due to the fact that increased size of the wafers and complexity of circuits being built have demanded larger pieces of equipment with much more exquisite control. To be sure, today's diffusion furnaces turn out ten times as many square inches of silicon as did the cheaper models of a decade ago, but the sales value of the output of a furnace has not increased.

In 1966, a small semiconductor company could be started with an initial investment of less than $1 million. Today, a similar company would require a minimum of $6 million. The design time and cost to bring today's state-of-the-art devices (16-bit microprocessors and 16,000-bit memory chips) is at least ten times, and probably 20 times what it was to bring state-of-the-art devices to market in 1966. Thus, the pendulum has swung in the sector of industry that provides sophisticated large-scale integration (LSI) devices for a mass market to favor the large and well-entrenched companies.

Semiconductor technology has brought a rapid succession of more and more able components costing continually less. The digital system designer at present has as basic building blocks the integrated microprocessor and memory chip. The potency of these blocks is illustrated by Table 1, which compares the parameters of a Fairchild F8 microcomputer, pictured without its power supply in Fig. 4, with the ENIAC, the first electronic digital computer. A brief study reveals that, for about $100, a system designer can now incorporate into *any* system he wishes a contemporary ENIAC, 20 times faster and 10,000 times more reliable, and requiring 56,000 times less power and 1/300,000 as much space. But better integrated circuit components of the same nature are ahead in the immediate future. Electron beam lithography and, ultimately, electron beam processing on the wafer promise still more dense integrated circuits. If the industry remains on the curve of complexity plotted as a function of time given by "Moore's law" (*2*), it will achieve a complexity of 10 million interconnected components by 1985. These developments constitute the on-going evolution in semiconductor technology. What is its goal? When it is reached, will the electronics revolution cease?

Table 1. Comparison of parameters of ENIAC with the Fairchild 8 (F8) microprocessor. Abbreviations: CPU, central processing unit; TTY, teletype terminal; ROM, read only memory; RAM, random access memory.

Item	Parameter	ENIAC	F8*	Comments
1	Size	3,000 cubic feet	0.011 cubic feet	300,000 times smaller
2	Power consumption	140 kilowatts	2.5 watts	56,000 times less power
3	ROM	16K bits (relays and switches)	16K bits	Equal amount
4	RAM	1K bits (flip-flop accumulators)	8K bits	Eight times more RAM in F8
5	Clock rate	100 kilohertz	2 megahertz	20 times faster clock rate with F8
6	Transistors or tubes	18,000 tubes	20,000 transistors	About the same
7	Resistors	70,000	None	F8 uses active devices as resistors
8	Capacitors	10,000	2	5,000 times less
9	Relays and switches	7,500	None	
10	Add time	200 μsec (12 digits)	150 μsec (8 digits)	About the same
11	Mean time to failure	Hours	Years	More than 10,000 times as reliable
12	Weight	30 tons	< 1 pound	

*For comparison in this table, the F8 microcomputer is assumed to consist of one 3850-CPU, one 3856-PSU (2K bytes of ROM), eight 1K RAM packages, and the static memory interface chip 3853, plus teletype terminal interface circuitry.

Let us look to the ultimate goal in the evolution of components. That goal is the availability of (i) microprocessors of arbitrary speed and reliability, (ii) memories of as large capacity and of as small size as one could want, and (iii) operational amplifiers with no limit of open-loop gain, response speed, input impedance, and output conductance. Of course, at the goal the cost of the components should not be a hindrance to their use. But this limit of component evolution is not the end of the electronics revolution—the revolution in electronic systems just then begins in earnest. The systems engineer is challenged to conceive and design systems that exercise the potency of the component. Programming the flow of information and control signals through the system becomes a problem focus. Thus, an impending center of action in electronics is conception and design of new systems which harness more potent components to tasks earlier adjudged interesting but impractical to perform.

Increased attention must be focused on the parts of the system, the transducers, to which the perfected electronics part interfaces. These transducers—sensors, actuators, display devices—have typically not benefited from the orders of magnitude of improvement that have occurred in the purely electronic devices. Transducers are susceptible to improvement through the use of strategies which succeeded in purely electronic components. Broad understanding of solid-state science suggests improvement in transducers; light-emitting diodes and photodetectors are existing cases in point. Microfabrication technique—photochemical machining, for instance—is readily extended to transducers. The transducer area will certainly become an important center of action in electronics.

Thus, we see two centers of action due to achieve increasing importance ahead, the areas of electronic systems and the component-like area of transducers. Both of these represent a wide range of opportunity. In each, there are unique opportunities for the new entrepreneur in operations of high mobility and modest size.

The Continued Importance of the Entrepreneurial Sector

We have seen in the recent history of semiconductor electronics an important role for new ventures. A major part of the success of the electronics revolution has come from companies that did not exist at the end of World War II. An account of the development of new companies on the San Francisco Peninsula has been given by Bylinsky (*3*). A recent report (*4*) of the Commerce Technical Advisory Board points out that young, high-technology companies exhibit larger growth in sales and in jobs (42.5

and 40.7 percent per year over the period 1969 to 1974) than do older companies classed as either "innovative companies" or "mature companies." In the period between 1945 and 1974, the group of six "mature companies" studied exhibited annual average growth in sales of 7.8 percent and in jobs of 1.9 percent. The five "innovative companies" studied showed average annual growth of sales and jobs, respectively, as 16.5 percent and 10.8 percent. Thus, the importance of successful new ventures is clear both for the economy and job market as well as for the provision of new electronic systems.

The need for and benefits of profitability in new ventures are sometimes poorly understood. Appropriate profit levels benefit all sectors. An incentive is provided to entrepreneurs for their initiative. To the company, profits provide capital for expansion or for new products. To the customer, a profitable enterprise is necessary to sustained production and continued service. To the government, tax revenue is provided, and to the public, products are provided concurrently with the provision of more jobs.

An Expanding Role for Schools of Engineering and Science

Semiconductor and computer electronics are founded on an increasing depth of scientific base. Accordingly, new professionals require an increasing level of preparation. This situation automatically involves the university in a stronger way than was the case a few decades back.

The university has long been recognized as an effective location for basic research. The coupling of the research mission with graduate education produces some essential economies. Educated people and research are coproducts in the university. Knowledge research by its nature is speculative; moreover, predictions of the nature, placement, and timing of breakthroughs are uncertain at best. These uncertainties are much less troublesome in the graduate school environment than in industry. A negative result may bear practically the same educational value as a positive one if it contributes to knowledge. This is an essential economy as compared to industry, where all bets are placed on the research result.

While the synergistic relationship between education and basic research (or pure knowledge research) has long been recognized, at Stanford University a strong synergism has emerged between graduate education and applied research in integrated circuits. A research program in integrated circuits demands an advanced working technology to be most effective. At the same time, problems of great contemporary interest to industry are hazardous to undertake in the university. The small university team is exposed to difficult competition from much larger industry teams. At Stanford, interesting applications not yet of commercial interest have motivated successful and demanding work in custom integrated circuits. A reading aid for the blind and a series of medical instruments requiring custom integrated circuits have provided dissertation topics for more than 60 Ph.D. candidates.

Many collateral positive effects have emerged. An important liaison with the medical community has been established in which medical doctors participate with engineers in the development of new instruments of great importance to their research or practice. These in-

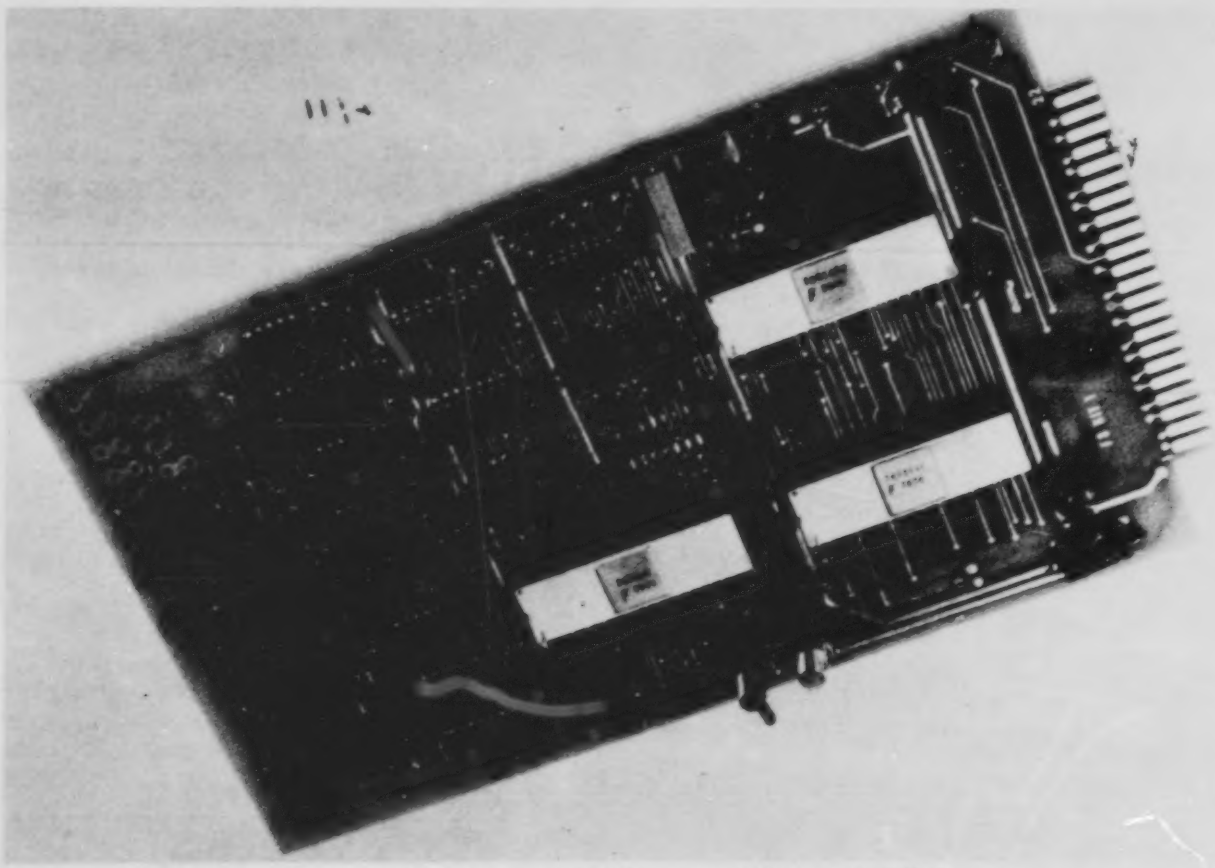


Fig. 4. The Fairchild F8 computer which is compared in Table 1 to the ENIAC.

struments (an ultrasonic imaging system, a series of blood-flow meters, new force transducers are examples) require custom integrated circuits which are the focus of study of graduate student teams. Liaison and collaboration with the integrated circuits industry has flourished. Industry has provided important inputs on contemporary technology. Early contact with promising candidates for employment has benefited both industry and the candidates. In one case, a new company was spawned to produce the blind reading aid using integrated circuits developed at Stanford. It has been profitably operating for 5 years with sales over the last year of $4 million.

From Our Crystal Ball

In the last quarter-century we have seen semiconductor electronics arise in the telecommunications industry and then provide the technological means to implement an emerging digital computer revolution, so that the individual today can personally own a more potent computational capability than his company or organization had then. Semiconductors have given the means to provide electrical power for space vehicles, to control their travel, and maintain their communication with the earth.

This electronics revolution has been fueled by a science base, promoted by government agencies vitally interested in the implements it promised, and carried out by new ventures with bold managers who perceived the opportunities and delivered outstanding performance. Electronics has grown in its utilization to address the urgent problems of defense and space, and its development from application to these problems has stimulated use in other areas. The $0.5-billion U.S. semiconductor business in 1960 was about half government and half private. By 1975, the $1.75-billion semiconductor industry had only a 22 percent government portion. We foresee that further development will proceed in the industrial and consumer sectors from the start already evident. We should like, however, to conclude with a forecast on the impact of electronics on the problem of this decade—the energy problem.

Semiconductor electronics will have a significant influence first in the area of energy conservation, in more efficient utilization of energy. Semiconductor high-energy ignition systems in vehicles are already improving efficiency. Microcomputer control of the variables of combustion (spark timing and fuel and air input) through measurement of engine temperatures and pressures and engine and vehicle speeds will make a further step in conserving fuel in transportation uses—uses that consume a quarter of our petroleum needs. In a similar way, more precise control of other systems that consume energy—heating, lighting, refrigeration, processing of food and material—can readily lead to economies not necessary with cheap fuel.

But, electronics also promises to bring power from renewable sources. We believe that photoelectric conversion of solar energy can be made viable as a source of power for terrestrial use within a decade. The harvesting of the average flow of 230 watts of solar power per square meter in the Southwest in significant quantities presents difficult problems. The necessary, large collection structures in a severe environment require a major capital outlay. The key to viability of solar energy is a sufficient conversion efficiency to support the capital outlay. An analysis (5) indicates that an overall conversion efficiency of 28 percent would produce power at the converter at 2 cents per kilowatt-hour. One method (5) with potential efficiency above this level is thermo-photo-voltaic (TVP) conversion. In TVP converters, silicon cells are irradiated by a 2000°K source heated by the sun. Silicon cells irradiated by a 2000°K source promise an efficiency greater than 40 percent if carrier lifetime can be kept as long as 1 millisecond. The processing of silicon cells at this point deteriorates the lifetime two orders of magnitude below the required value. We predict that perfection of photovoltaic converters can occur within a decade if sufficient effort is mounted to address the problem.

To us, this seems less remote than did the construction of components for Minuteman in 1958. At that point, a failure rate of 0.0007 percent per 1000 hours was required with a then-current failure rate for transistors of about 1 percent per 1000 hours. Actually, improvements were achieved in about 3 years and these reduced failure rates to about 0.0003 percent per 1000 hours.

The solar conversion program must be addressed with the same composite of forces that brought reliable components to Minuteman. Outstanding teams of engineers and scientists must attack the problems. The government agencies involved must exhibit the same vision of mission and persistent support of goal attainment. As viable designs emerge, production and installation must proceed with the same industrial competitiveness and efficiency that has characterized the past three decades of the electronics revolution.

References and Notes

1. W. Shockley, *Electrons and Holes in Semiconductors* (Van Nostrand, New York, 1950).
2. G. E. Moore, "Progress in digital electronics," paper 1.3 presented at the International Electron Devices meeting, December 1975.
3. G. Bylinsky, *Fortune* **89** (6), 128 (June 1974).
4. R. S. Morse, "The role of new technical enterprises in the U.S. economy," Department of Commerce report, January 1976.
5. R. Swanson, personal communication. This analysis assumed a plant cost of $100 per square meter.

6. The authors dedicate this article to William Shockley to acknowledge his early and monumental contributions on which the electronics revolution is based. The contribution of J.G.L. to this article was supported by grant Eng-75-22329 from the National Science Foundation.

Computers in Banking and Marketing

The technology is here, with promises and problems for both consumers and corporations.

F. E. Balderston, James M. Carman, Austin C. Hoggatt

Businesses and consumers in the financial industries and in retail trade engage in varied sets of transaction activities. Computers are modifying and will modify the organization of these activities. Just as in previous decades the telephone and the automobile changed consumer opportunity and patterns of search, so too may the computer. The automobile, for example, stimulated suburbanization and the subsequent concentration of large-scale shopping centers in outlying areas where car-borne consumers could buy a large variety of convenience and specialty goods at one time and could also visit a number of stores for comparison shopping. We will discuss computer applications that relate consumers to firms in financial industries and retailing, and we will look to major future trends in services.

Our subject matter includes (i) on-line service to depositors, (ii) electronic funds transfer systems, (iii) inquiry and verification systems for credit card transactions, (iv) reservation systems, and (v) point-of-sale electronic registers.

These applications have in common the property that they support activities at the point of sale, the point at which the business establishment and the consumer come together to engage in a transaction. This focus on the point of sale will, in turn, require us to take note of consumer attitudes and concerns as they may affect the design of computer-supported systems; many other types of business computing require no such sensitivity to the marketing interface.

Consumers may have much to say about which of these systems will endure; they may endorse and patronize the systems that give better service or reduce consumer costs, and they may resist and defeat systems that are irritating or too costly. Business organizations could probe the potential consumer response patterns by developing and offering new point-of-sale systems but they will make these investments with great uncertainty. It is nearly impossible to predict the outcome of this process.

The authors are faculty members of the Schools of Business Administration, University of California, Berkeley 94720. They shared equally in the preparation of this article.

Computers in Banks and Other Financial Firms

We will discuss five types of applications: (i) batch systems, (ii) on-line inquiry and teller assistance from terminals, (iii) automatic teller machines, (iv) electronic funds transfer systems, and (v) other financial services that are supported by computers.

Batch systems. In the sequence of interactions between the customer and the financial firm (commercial bank or savings institution), we distinguish conceptually between the stages of inquiry and verification and the stages of execution and recording of financial transactions. Many commercial banks, for example, use batch processing of each day's volume of checks and deposits to produce an updated listing during the night of each depositor's account status for the next day's banking activities. Tellers consult this listing during the day to determine whether withdrawals requested are within the balance limit and to answer customer requests for information on the status of their accounts.

Many commercial banks have progressed, since the late 1960's, to the point of operating on a batch basis several high-volume processes by means of multiprogramming. These may include not only the daily check processing and updating just discussed, but also preparation of the bank's trial balance and ledger accounts.

The large-scale, batch-process systems typically centralize account files of multibranch institutions and have been implemented with complex programming. An arrangement in New York City permits a number of commercial banks to settle financial claims with each other by computer. At the end of each working day, final balances are communicated to the regional Federal Reserve Bank for recording there, but the law still requires paper documentation to follow.

There has been success, however, in standardizing input requirements so that preauthorized payments (payrolls of large employers; entitlement payments of the Social Security Administration, and so forth) never enter the "paper" system: they are prepared on magnetic tape, and the bank credits the payment to each account by passing the tape input through special programs. These preauthorized, paperless payments, which have operated for several years, have grown to account for approximately one-tenth of the Social Security payment items, and they are helpful both to the payee in assuring prompt receipt of the payment due and also to the payment source by reducing claims for lost payments. These payment systems do reduce the "float" (several days of delay before a check clears) available to the payer's bank, and the bank cannot, therefore, reward the payer as it would if the payments stream occurred with the usual distribution of delays, resulting in the maintenance of the payer's account balance at a higher level for several days after each high-volume payment cycle.

On-line inquiry and teller assistance from terminals. On-line teller assistance terminals now used by numerous savings institutions illustrate how the early part of the transaction sequence may be handled. The teller's terminal can address the customer's account in a central computer memory. All transactions are stored until the account holder's passbook or statement has been updated on the terminal printer. The teller has a limited series of available commands, each actuated by a single key. The terminal has provisions for on-line updating of any high-volume customer file, including savings accounts, collection accounts, passbook loans, and real estate loans. For any of these accounts it can deal with withdrawals or deposits, updating of interest and (if necessary) calculation of penalty interest, record changes to the master file, and teller totals for every transaction. The teller can also, by command, receive account status information on cathode-ray tube or printout before proceeding with any of the above operations. The system has printer capability so that the teller can audit the input information for accuracy before giving the command for the next steps of processing. The same system contains the

← *Courtesy Communications Satellite Corporation (COMSAT)*

general ledger of the institution, although hardware and programming controls are intended to restrict access to this system to authorized accounting personnel. For handling the institution's payable accounts, the authorized person can enter the appropriate name and accounting information and have a payment check printed on-line, with the updating of the ledger account also performed on-line.

In one well-established version of this type of system, several major operations are still batch processed overnight after each day's transactions: the transfers from savings account files into ledger accounts, the calculation of the daily trial balance of the institution, and the cross-checks and verification of accounts. The on-line system for one fairly large institution includes 400 terminals, most of which are in the branches that handle customer savings transactions; the institution's loan departments have their own display and transaction terminals, as do the general administrative departments of the headquarters. This large network of terminals is serviced by leased lines to a central processing unit operated by a specialized company with which the institution has a contract. Minicomputers scattered throughout the terminal network control the authorized access to different account files and provide message editing, buffer storage, and message concentration between the terminals and the main central processing unit (CPU) and storage.

Automatic teller machines and credit card systems. Automatic teller machines (ATM's) for commercial banking being developed are in various stages of pilot testing. The most inclusive concept of an ATM would be a machine the customer could use to accomplish any business that could now be done through a human teller. So far, the machines being tested are more limited.

One such machine is the cash dispensing machine. Chemical Bank of New York, among others, has installed these at street-side locations. The customer inserts a card, gives a password, asks for an amount of currency, and receives it at a slot.

Other ATM's are being designed to be used within a bank's branch location, either (i) as teller-assistance terminals to make deposits and withdrawals or to provide information concerning account status or (ii) to enable the customer to perform these functions alone with appeal to a bank employee if the operating instructions are too difficult to follow. It is much easier to design the terminal system if it is intended to perform only as teller assistance.

The ATM's designed to be placed at retail locations, with no bank personnel present, would offer the retail merchant a way to authorize credit or the consumer a way to acquire cash other than through the store's credit department.

Credit card transactions may now have automated support as well. One limited but useful function is that of verification and authorization of the card that is presented. At many retail establishments, this must now be done by telephone inquiry if the retailer does not wish to take the risk of defaulted payment. One system provides for negative verification of the card in the sense that its number does not show up on a list of "bad cards."

BankAmericard Service Corporation has a switching service for its member banks for card verification. An extension of the system, now being tested, allows the merchant to obtain verification either by phone call or by terminal inquiry. The merchant puts the customer's card and the sales draft into an electronic device. Then, the transaction amount is entered from the keyboard. The merchant's identification is picked up from the terminal's permanent storage. The message goes to the bank's computer. The originating bank that issued the credit card is then asked whether the card is good for the amount in question.

For inquiry and authorization as well as, potentially, for updating and transaction messages, terminal networks driven by minicomputer centers are being developed (*1*). The Bank of America's development will service, by means of 6000 terminals, inquiries concerning the status of accounts coming from its 1000 branches in California. This inquiry and message system is just one module of a series of modular arrays. Each will consist of a specialized machine configuration and operating system, designed for one set of purposes. This strategy is an alternative to that of adding more services and interactive tasks to an existing centralized and multiprogrammed system. It is not yet clear which of the two approaches is likely to win out in large financial institutions in which the capital and organizational choices are made centrally.

Electronic funds transfer systems (EFTS). The EFTS has long been discussed as a basis for the checkless and cashless society; development and pilot tests are now under way. To realize EFTS would require: (i) a terminal at each point of sale, which could not only request verification and authorization to debit the customer, but which could then begin to execute the transaction, crediting the retailer and debiting the consumer's deposit account; and (ii) a switching network to link all the financial institutions whose accounts would be affected by the flow of transactions. Such systems arouse important issues of public policy, such as privacy and security protection. Apart from these issues, widespread adoption of EFTS would affect the market structures of the financial industries and would tend to change the boundaries between retail trade and the banking system.

One basic question concerns the legal status of a payments terminal. Is it, in itself, a "branch" of the bank or savings and loan association to which it is connected? Much legal and economic analysis of EFTS will be required before this and related questions are resolved (*2*).

Widespread adoption of EFTS would, many believe, decrease the importance of branch offices of banks and savings and loan associations, which have provided services at convenient locations to their consumer customers; excess capacity and branch closings may well occur. Also, it can be presumed that the size of the capital investment required for EFTS will be discouraging to small banks and savings and loan associations and that some of them will not survive a large transformation in financial transactions systems. Some also believe that EFTS will further help the consumer allocate financial resources advantageously among types of accounts (for example, time deposits versus demand accounts), types of institutions (savings and loan associations versus banks) and ways of holding assets (treasury and municipal bonds versus accounts in financial institutions). The resulting increase in the fluidity and volatility of financial capital may well raise new problems of monetary control. Citibank, which has made a major developmental investment in the new technology, does not apparently believe that monetary instabilities will result (*3*).

The competitive equilibrium between banks and savings and loan associations may be modified by these changes in the technology of payments, reducing to some extent the importance of mandated interest-rate differentials (*4*). Many of the thorny public policy questions caused by EFTS are now being studied by a national commission on EFTS.

It is reasonable to predict, however, that competition among types of automated services will rage for some time, that competing EFTS mechanisms and switching systems will battle it out, and that, for many years to come, the present

paper-based payment systems (checks and credit card sales drafts) will continue in heavy use, coexisting with EFTS.

All of the schemes for EFTS that we have surveyed have the same fundamental cost structure—costs are largely fixed with only small variations in the cost of operation until capacity is nearly reached. At some point near capacity, these systems will display response degradation just as all other computer systems do. Unless new methods to maintain performance and security are developed, new forms of regulation will surely emerge. Overlaying such controls on the already stringently controlled banking system could produce interesting regulatory problems stretching into the indefinite future. A scheme to naturally limit growth of networks would be a valuable social asset.

Computers in Marketing

In this section we discuss the present and projected state of application of computers in the marketing of goods and services outside of banking. Discussed are applications in reservation services and retail trade. Not discussed in detail are four areas that we feel are necessary for a complete overview of our topic: computers in insurance, in securities markets, as consumer products, and in distribution of goods.

In insurance, computers are used to write new policies and process claims. This means that actuaries now have very current data on what populations are buying a new policy and how many claims are being filed on these policies. Thus, the actuary can now continuously monitor each type of insurance policy and the claims against it and detect and correct pricing mistakes much more rapidly.

The automated securities market raises an entirely new set of issues regarding exchange transactions. The goal now is to tie all regional markets into a system that functions as a single network, but another goal that may be technically possible is to replace all traders with a computer. Some research has been completed on the role of the market specialist in such a system, but many more intriguing and important research questions concerning public policy remain.

Similarly, the computer as a consumer product offers a yet untapped mine of ideas and problems. The pocket calculator is the standard example of such a product, but there are others. The electronic automobile diagnostic center, for example, is a service that has increased efficiency in a business field that has seemed in danger of being overtaxed. In contrast, the home video game and its future extensions may divert a major part of the problem-solving activity of the country to synthetic and perhaps irrelevant problems.

There have been computer systems for business purchasing and inventory management since the earliest applications of computers to business. Virtually all sellers in the industry offer some equipment and software to serve these needs. The major changes being made at this time are caused by the general availability of cathode-ray tubes and data communication networks.

Computerized reservation services. Remote terminals tied to a large CPU provided a basis some years ago for reservation services in airlines, hotels, automobile rental, and spectator entertainment events. These systems have been around long enough that it is now possible to develop some insights about their impact on the marketing of these services.

First, they are obviously efficient and powerful. Airline systems have provided more efficient aircraft use, better customer service, and a comfortable way to meet federal restrictions on the overselling of space. Still, these systems appear to have unused capacity. Most airline computers will now add seat selection, car rental, and hotel reservations to the information they store and transmit. Specialized printers at terminal sites now write tickets so efficiently that they have replaced what has historically been a major inefficiency affecting both the buyer and the seller. The end of the benefits for the traveler is still not in sight. For example, one hotel had an on-line system to keep accounts of charges incurred by its guests; it used this system to provide a wake-up service, thus relieving switchboards of a time-consuming daily task. The computer calls the room at 10-minute intervals until the phone is answered.

Buyers of hotel and airline systems now have multiple vendors to select from. National Cash Register's "Inn-Tak" is a typical modular system for hotels and motels with more than 500 rooms. The innkeeper can buy either the reservation module, the guest-accounting module, or both. Each costs approximately $60,000.

Second, some car reservation services provide insights into the demand-stimulating value of computer systems. The computer has all the paper work completed when the traveler arrives at the counter. The customer need not even produce a driver's license, because the computer already has that information. The "instant service" is a way to hold customers.

Hotel and motel reservation services provide an interesting example of another kind of promotional and competitive impact of computer reservation systems. Hotels have found that reservation services help to build customer loyalty. They have therefore formed—by integration, franchising, or cooperative arrangements—chains of hotels or motels to provide service to the traveler at every stop on his or her trip. These chain organizations use either their own in-house reservation systems or contract reservation services. Multifirm reservation services like the American Express "Space Bank" are no longer advertised. There are, in fact, only three "cooperative" reservation services left. Two of these are supplementary to systems already operated by major motel chains for their own guest reservations. These three service companies do not advertise to the public at all. The individual motel or travel agent simply advertises that it can provide reservation services to individual customers.

Computers in retail trade. Perhaps the greatest impact of computers in marketing in the 1970's is the point of sale (POS) terminal. These computers are exciting because of their potential for efficiency in retail operation, because of their potential for still further innovation, and because they bring the customer so directly into contact with the computer.

The simplest innovation a retailer can make in this area is simply to replace his mechanical cash registers with electronic ones. There are about seven suppliers in this field, with National Cash Register having about half of the market. Sales of these registers now total more than $800 million per year and are growing at an annual rate of about 28 percent. The electronic register that requires no connection to the CPU costs less than mechanical registers, reduces cashier training time by as much as 80 percent, processes a sale in one-half to two-thirds the time of old registers, and records more information concerning transactions. Thus, even the small retailer is going electronic.

The next step is to convert the POS register to a terminal tied to other terminals. In the days of the large CPU's, each POS terminal transmitted all information to a central computer, usually not even in the store. The present design concept, however, is to process the data on minicomputers at the level in the orga-

nization where the data are to be used. Cash payments are completed at the register; departmental sales analysis is done at the branch store; only summary information leaves the branch store. This practice, its advocates claim, provides more timely information where it is needed and reduces data transmission by 20 to 40 percent. Decentralization of computing also has an advantage over giant centralized systems with regard to data security. A fire will not destroy all the records of a firm; downed telephone lines will not put a store out of business for hours or days. However, the advocates of centralized data processing have not surrendered yet.

The POS terminals can next be tied to credit verification centers. The store that accepts both company and bank or travel cards might be tied into a number of credit verification centers, some in the store and some hundreds of miles away. If EFTS becomes a reality, the POS terminal stands ready to deduct your bill at the store from your bank account and add it to that of the store.

The automated service station is an example of credit verification and self-service both controlled by a minicomputer. In a typical configuration, the terminal will accept $1 bills or a company credit card. When the credit card is inserted into the machine, it triggers a phone inquiry to a distant credit verification center. If the card has been reported stolen, the terminal will capture the card and not return it. If credit is approved, the terminal will instruct the customer on how to pump his own gasoline. Of course, the terminal will also provide a summary of all transactions at the end of each working day.

The next addition that the POS terminal can accept is the laser optical scanner that can replace manual entering of information. The problems here are twofold: universal product coding (UPC) and consumer acceptance. The grocery industry worked for many years to achieve UPC. Even without optical scanning, UPC has increased efficiency. The achievement of UPC was a major accomplishment for cooperation within the industry between manufacturers and distributors.

For nonfoods, the problem is more complex. The suppliers are less consolidated and the assortment of products is much broader. Also, retailers of nonfoods realize that characters must be readable by humans as well as machines. Most items will have to be marked at the store with both code number and price in a system that has optical character recognition (OCR). A few OCR systems are in operation today, but there is a catch. Item marking plus OCR scanning for the wide variety of goods that department stores must handle pushes their cost to perhaps 500 percent of the cost to the grocery store. It is doubtful that such a system can be justified in terms of the costs, at least in stores where check-out queues are not a problem.

But even in grocery stores, optical scanning is not sweeping the country. Scanners are now found experimentally in perhaps 100 stores throughout the nation. These tests show productivity gains of about 25 percent at checkout. However, if item pricing is required, about one-third of the cost savings disappears.

We have now developed the POS terminal with the capabilities to analyze sales, to control inventories, to transfer funds, and to optically scan data. Perhaps we can replace the whole store with such a device. But we are probably not very close to the day when two-way closed-circuit cable television tied to the POS terminal will replace the store. Catalog selling by means of the printed book and the telephone is growing and is very efficient. More important, it is not clear that consumers will stand for more automation in the name of efficiency alone.

Assets and liabilities of the use of computers in promoting retail sales. The history so far of computers in direct contact with consumers provides some lessons on just when computers have positive promotional value and when they have negative promotional value. In addition, we suggest that there may be a third category in which consumers are neutral about computing in the short run but may become negative in the long run.

Computers can have positive promotional value when they yield an obvious consumer advantage. In this case, sellers are well advised to make promotional hay while the sun shines. Avis Wizard is a good example of a way the computer can improve service.

Optical scanning in food stores provides the classical example of negative consumer response. Consumers like the additional speed at the check-out counter, but they are not willing to give up item pricing. Consumers do not trust supermarket pricing anyway. Removing prices from items in the name of efficiency has not met with enthusiasm. In fact, the legislators of four states (California, Connecticut, Massachusetts, and Rhode Island) quickly passed bills prohibiting the removal of item prices.

Some other new technologies may or may not encounter long-term consumer resistance. Computerized banking systems could be simpler if consumers did not insist on the return of their canceled checks; but, so far, consumers want to keep this form of personal record for their own audit control. Fearing that consumers would refuse to give up possession of the printed account record, savings and loan associations also seem to be hesitating to eliminate the savings account passbook. Besides resisting loss of control over their own paper records, consumers may not adapt willingly to changes in the timing of payments. For instance, checks now require a few days to clear, and many credit card transactions are not settled for several weeks. Will consumers resent compression of these payment cycles?

Consumers are likely to be either indifferent to the new systems or mildly hostile to them as one more new, depersonalized system. Why should consumers be interested in paying for the new payment services? If consumers do not care enough to finance the new schemes, this creates a struggle between the retailers and the financial institutions over who will bear what costs.

Some Possible Social Responses to Pervasive Computer Systems

The growth of information networks with memories raises issues of security and individual privacy. Large investments may be made to protect the privacy of individuals and the integrity of systems from breaches of security, but both theory and practice tell us that such investments can, at best, reduce but not eliminate the probability of harm. At some point, the marginal investment cost will exceed the value of the marginal improvement in security. These large-scale interconnected financial and marketing systems, since their control programs are merely information having special meaning to a computer, will be vulnerable to the terrorist or thief.

Realizing that complete security is not possible, we could perform an interesting social experiment. Suppose two electronic funds transfer systems were offered with identical financial systems capabilities. System 1 is very secure and devotes much of its effort to protecting the privacy of customers; this expense is passed through in-service charges to the users. System 2 has easy access for people or organizations who may use the data in the files for any purposes they wish (since the blackmail potential for this information is nil, presumably the users would be doing marketing research). Customers are paid a premium for using system 2 and the commercial

interests pay the costs of operation. This scheme has the interesting prospect of allowing a market choice for consumers with regard to privacy. It also has the potential for cost reduction, as many a type 1 person could be transformed into a type 2 person when the costs were made explicit.

The quality of service is also important. While information is discrete, complex systems that handle and deliver information are subject to response degradation ranging from minor response delay to information "brownout" or even collapse. There is also a risk of irrecoverable losses of stored information.

Automobile recalls for safety, environmental impact, and consumer satisfaction are good examples of the far-reaching changes which are in store. In such systems statistical quality control is not enough. Records must be kept initially on the product identity and on the seller's identity and location. Records are also kept over time; as defects become apparent in use, the consequences are determined, and some subset of already sold items is called in for warranty repair. Indeed, with the built-in computer monitoring systems necessary to meet the stringent environment controls of 1980, it is possible that the decision to recall could, for some conditions, reside in the control computer of the individual car and be evoked when failure was detected in use.

This raises an interesting possibility for decentralization. As miniaturization continues and more powerful devices with memory become available, it will be possible to place monitors in all sorts of equipment (home appliances, and so forth) and to keep a log of activity. A record of demands and performance available to manufacturers and consumers would be a useful basis for warranty decisions, and it might relieve a burden of centralized data storage and retrieval.

A common thread in these examples is growth toward more complexity. We have mechanisms at hand that permit great expansion of bureaucratic regulation (public and private) and that may induce us to choose organizational forms of increasing complexity. We should be cautious about embracing these possibilities. Automobile recall requires considerable information. A matrix of interactions between dealers and owners extends over time and requires storage that increases at least linearly with time. Cost accounting for such information systems is not yet well developed, and storage without retrieval has been with us ever since man invented librarians, so it is unlikely that these systems will be well run.

Random isolated failures of complex information systems are expected and accepted. The social consequences of the general failure of such systems are less acceptable and less understood.

Any EFTS should be designed to be robust against general system collapse. Experiments with parallel systems should include careful tracing of costs and benefits. These systems should not be imposed on society before they have been thoroughly tested.

References

1. J. D. Foster, *Datamation* 22 (No. 7), 54 (1976).
2. W. Baxter, P. Cootner and K. E. Scott, *Retail Banking in the Electronic Age: The Law and Economics of Electronic Funds Transfer* (Allanheld Osmun, Montclair, N. J., 1977).
3. CitiBank monthly economic letter (December 1976).
4. K. J. Thygerson and K. Reich, paper presented at the 2nd Annual Conference of the Federal Home Loan Bank, San Francisco, Calif., 10 December 1976.

The Impact of Integrated Electronics in Medicine

Robert L. White and James D. Meindl

Health care presents a singular opportunity to improve the quality of life in our society through electronics. In modern medical practice the premier diagnostic and monitoring tools are electronic instruments as illustrated by the x-ray and electrocardiograph. An electronic prosthesis, the implantable cardiac pacemaker, is essential to life for many patients. In medical research, a myriad of electronic instruments ranging from electron microscopes to pinpoint-size microelectrodes contribute enormously to the prevention, identification, and treatment of disease.

During the past 16 years the monolithic integrated circuit has had a revolutionary impact on electronics. Throughout this period the number of elements per chip, or the functional complexity of integrated circuits, has doubled annually. Today 64,000-element chips are commercially available. Because of their low cost, excellent performance, high reliability, and small size, integrated circuits have made possible whole new industries, and their impact pervades our lives. In particular, the unique capabilities of integrated electronics can be applied to medical instruments with enormous benefit to the quality and availability of health care. The purpose of this article is to illustrate through a number of examples how health care has been expanded through the new diagnostic, monitoring, and prosthetic instruments now possible through integrated electronics.

Diagnostic and Monitoring Instruments

The ideal diagnostic instrument provides definitive data on the patient's condition, causes him no harm or discomfort, and is convenient, reliable, and economical for the physician or his medical associates to operate. A monitoring instrument imposes the additional stringent requirement of virtually total freedom from the need for a human operator. Because of their noninvasive character, transcutaneous instruments such as computer-aided tomographic x-ray scanners, electrocardiographic monitoring units with automatic arrhythmia detection, and ultrasonic imaging and blood flowmeter systems offer the greatest possibilities of fulfilling these idealized specifications. However, in view of the limita-

Dr. White is a professor of electrical engineering and materials science and is director of the Institute for Electronics in Medicine in the School of Engineering, Stanford University, Stanford, California 94305. Dr. Meindl is a professor of electrical engineering and director of the Stanford Electronics Laboratories in the Stanford School of Engineering, Stanford University.

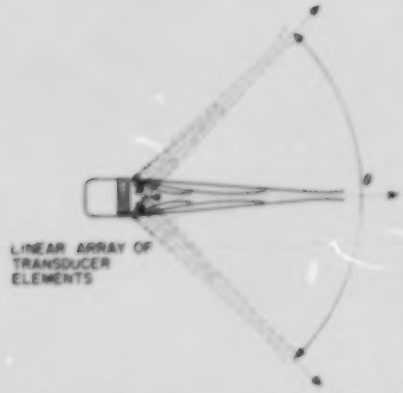


Fig. 1. Electronically scanned and focused piezoelectric transducer array.

tions of noninvasive techniques in medical research involving animal models of human disease, invasive instruments such as subcutaneously implanted telemetry systems for measurement of blood flow and pressure are invaluable in their ability to collect physiological data that would otherwise be unavailable. We now review several examples of both noninvasive and invasive diagnostic and monitoring instruments, with emphasis on capabilities possible only through integrated electronics.

Computer-aided x-ray tomography. In 1971, when a British company with no prior position in x-ray equipment (*1*) introduced the tomographic head scanner, the long-stagnant field of x-ray diagnostics was revolutionized. The order backlog for this scanner, and its instant competition, ran into the hundreds of units, although the instruments cost from $300,000 to $600,000 each. The new instrument (*2*) relies on a different approach from the conventional x-ray shadowgraph and gained instant acceptance because of its capability for resolving tissue densities on the order of ½ percent, as compared to the 25 percent difference resolvable by conventional techniques. Such resolution allows the location and identification of tumors as well as the distinction of body fluids from soft tissue, both difficult-to-impossible by shadowgraph techniques. The tomographic scanner operates by recording the attenuation of a narrow x-ray beam as it scans the patient from all angles in a single plane. Each element of tissue in this cross section is therefore observed from a number of directions, and it is possible, by computer processing of the totality of the recorded data, to separate the absorption coefficients, accurate to ½ percent for each small element. The absorption data is then converted to a brightness scale, and a tissue density map of the object plane is plotted. In this diagnostic technique, integrated electronics is used throughout the digital data processing system, without which computer-aided tomography would not exist. Indeed, if one takes the extreme view

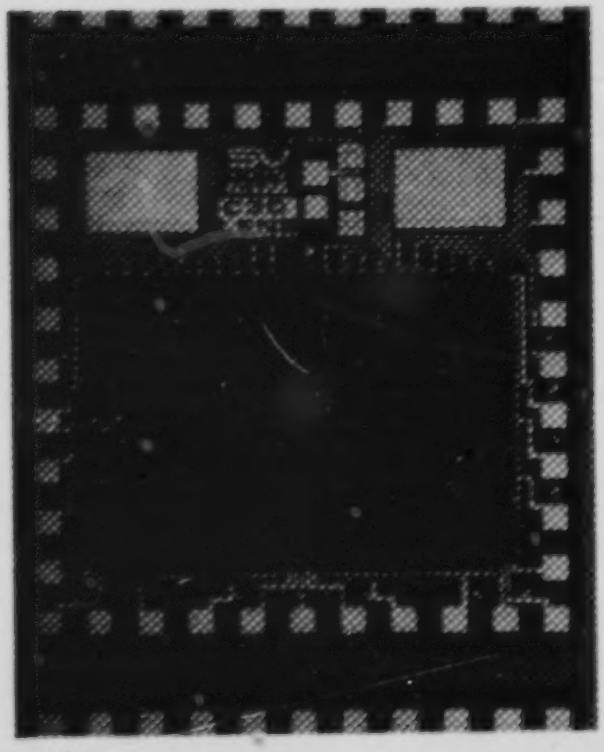


Fig. 2. Cascade charge coupled device (C3D) photomicrograph. Chip size is 3 by 3 mm.

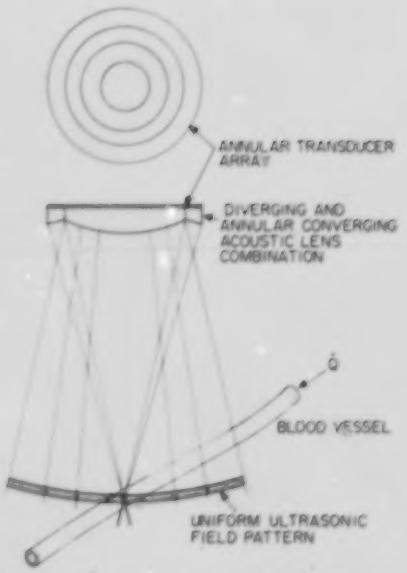


Fig. 3. Noninvasive volume blood flowmeter.

that computers have their present power, price, and pervasiveness only because of the advent of integrated circuits, one can argue that medical practice has been totally infiltrated by integrated electronics from appointment call to billing.

Coronary care heart monitors. It is imperative that cardiac performance be monitored continuously for patients recovering from a heart malfunction. To have a nurse or technician continuously watching an electrocardiograph display is both unacceptably expensive and insufficiently reliable, since the significant warning events are occasional irregular beats (arrhythmias) out of hundreds or thousands of regular beats. The solution has been provided by computer monitoring of the patient's electrocardiogram (*3*). The computer can be programmed to analyze each heartbeat in a multivariate fashion, identifying such features as the height and duration of the principal electrical event (the QRS complex) and the interval between beats, and on the basis of these data classify each beat as normal or abnormal. The computer then displays at a central monitoring station of the coronary care unit such information as heartbeat rate and the number of abnormal events occurring in a given interval, perhaps sounding an alarm if abnormal events exceed a predetermined rate. Such monitoring systems, which provide patient care more reliable than a human observer, are possible only because of the computational power provided by integrated circuits.

Ultrasonic imaging systems. In comparison to x-rays, ultrasound is a recent

noninvasive diagnostic modality that offers superior characteristics in some circumstances and inferior ones in others. Applications in which ultrasound frequently provides advantages include (i) imaging moving targets, such as the heart or the fetus, (ii) imaging the pregnant uterus where minimal x-ray exposure is advisable, and (iii) imaging soft tissues such as muscles and blood, with very similar densities but very different acoustic properties. Typical applications in which x-rays are often more useful include imaging the lungs or the brain where air passages and the skull, respectively, severely attenuate ultrasound.

The basic principle of operation of most ultrasonic imaging systems is similar to that of sonar or radar. A short pulse of high frequency 1.0 to 5.0 megahertz ultrasound is emitted by a piezoelectric transducer contained in a handheld probe, in contact with the patient's skin. As the packet of ultrasonic energy propagates through the body, most of it is lost through tissue absorption; but a small amount is reflected at boundaries between tissues of differing acoustic impedance. It is the reflected component, which returns to the transducer, that provides useful information. Until recently, images were generated from these reflections by manually scanning a single transducer element attached to a mechanical arm equipped with several potentiometers providing electrical signals describing probe position and orientation. Feeding these potentiometer signals along with the returning echo signals to a cathode-ray tube display produces a cross-sectional image. Typically, such images are low in resolution as a result of the use of a single, large-diameter transducer element emitting a collimated acoustic beam. In addition, manual scanning precludes real-time imaging of moving targets, such as the heart. Integrated electronics can be used to overcome these two limitations and compose echoes for image formation (Fig. 1). All transducer elements in the linear array are used in a controlled time sequence for both electronic focusing and electronic scanning in a sector pattern to produce a real-time cross-sectional image. The special features of this approach are high resolution at virtually all ranges, since "electronic zoom" is used dynamically to adjust the focus to each depth, and, in addition, a large sector angle, $\theta = \pm 35°$, provides a panoramic field of view. This scheme is particularly suited to high-resolution dynamic cardiac and abdominal imaging.

In addition to the transducer array, integrated electronics provides several other key contributions to this system. Each transducer element must have an associated transmitter circuit, and these must be triggered in a carefully timed burst to permit launching an acoustic wave in any specified direction within the $\pm$ 35° sector. A carefully matched set of preamplifiers must be connected to the transducer array to receive returning echoes. The output signals from these preamplifiers are fed to a corresponding set of electronically variable time delay lines which form the heart of this system. In essence, focusing on a particular point in the target sector consists of equalizing the time delays from that point to the common output node (or summation point) of the delay lines. That is, differences in acoustic path delays from a particular target point to various transducer elements must be compensated precisely by a complementary set of electronic time delays preceding signal summation.

This highly specialized requirement of medical imaging systems has been met by a unique integrated circuit. It is a cascade charge coupled device (C3D) which consists of a monolithic set of variable delay lines whose time delay is the product of the number of stages (or length) and the electronically controlled transfer rate of a signal charge packet from stage to stage (*4*). A photomicrograph of this C3D is shown in Fig. 2. The section of the chip labeled f_1 accepts a separate preamplifier output for each of its 20 parallel delay lines. The sections f_4 and f_5 correspond to acoustic (or optical) prisms whose two acute angles, or height, are electronically variable by changing the C3D transfer rates which control the scan angle of the imaging system. Sections f_2 and f_3 correspond to an acoustic (or optical) lens doublet whose radii of curvature are electronically variable by changing the transfer rates, which control the range of focus of the imaging system. The output of section f_1 represents the signal summation point. This remarkable integrated circuit zoom lens provides the basis for extremely high performance ultrasonic imaging systems in the future.

Ultrasonic blood flowmeters. The capability to perform harmless, repeated, accurate, noninvasive measurements of volume flow is necessary in the diagnosis and treatment of countless medical prob-

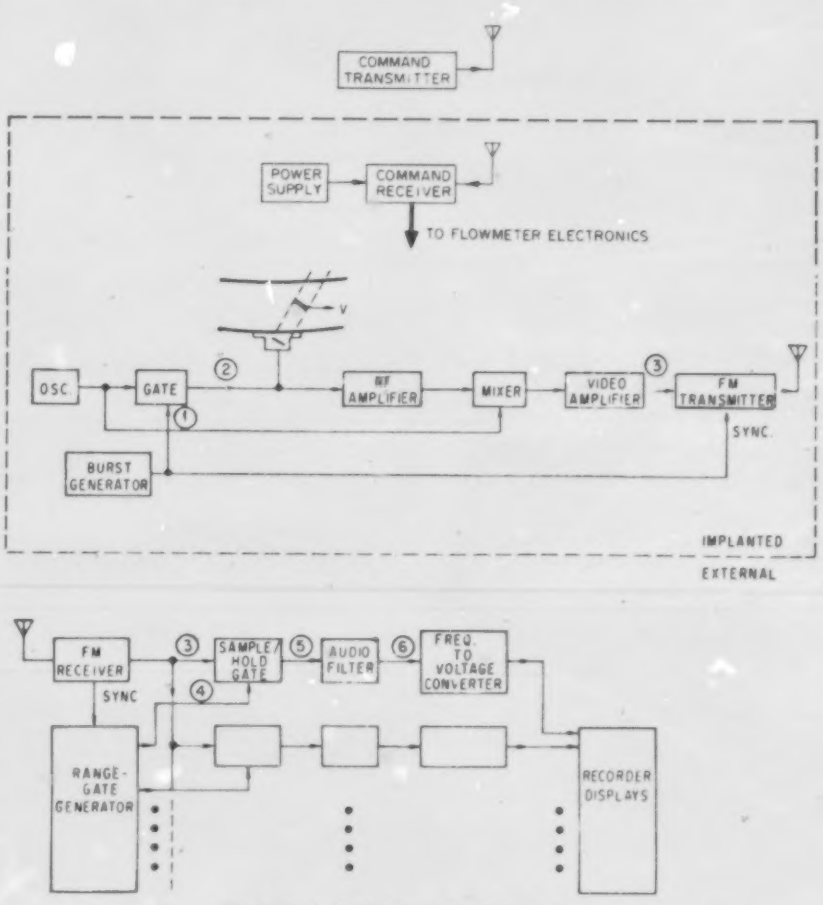


Fig. 4. Implantable flowmeter block diagram.

41

Fig. 5. The Optacon in use.

lems. A new concept for transcutaneous measurement of volume flow which utilizes an integrated annular array of piezoelectric transducers is illustrated in Fig. 3 (5). In this scheme, a short burst of high-frequency excitation is applied to the three inner elements of the array at three unequal amplitude levels, so that in combination with the diverging acoustic lens, the transducers generate a uniform ultrasonic field pattern within a given range interval. This interval must include the entire lumen of the blood vessel in which flow is to be measured. Following the transmit pulse, the three inner elements are connected in parallel to receive returning ultrasonic energy which is backscattered and Doppler-shifted in frequency an amount Δf by the moving red cells in the blood vessel. This Doppler frequency shift is directly proportional to the velocity of blood flow. In addition, the outer annular transducer is used to receive returning ultrasonic energy which emanates from only a portion of the lumen (that is, the "small area") as defined by the annular converging lens.

To convey the fundamental principle of operation, it is convenient to assume that the "small area" is known as a result of the characteristics of the converging lens. Hence, by electronically forming a ratio of the Doppler signal power returning to the three inner transducers to the Doppler signal power returning to the outer transducer, the area of the lumen is determined. Furthermore, by computing the centroid of the power spectrum of the Doppler signals returning to the three inner transducers, in essence, one can determine the average velocity of blood flow in the lumen. The product of lumen area and average velocity is volume flow Q.

Careful analysis of this approach indicates that the instantaneous volume flow measurement is independent of the angle of orientation of the transducer relative to the vessel, the blood velocity profile within the lumen, and the shape of the lumen (5). Extensive medical applications of this new system are anticipated.

Implantable telemetry systems. The ability to prevent and treat disease in man has depended on animal research. In cardiovascular research, for example, it is now widely recognized that acute physiological measurements in anesthetized, open-chest animals often cannot be duplicated in intact awake animals. It follows that concepts derived from such experiments are not necessarily applicable to intact unanesthetized man. Therein lies the rationale for implantable telemetry systems in animal research. These unique instruments simply provide the means for continuous physiological measurements that otherwise cannot be achieved. Because of the complex functional capability, small size, high reliability, and low power drain required of implantable units, integrated electronics must be used.

Two of the most useful measurements that can be made in cardiovascular research are blood flow and pressure. A block diagram of an implantable, ultrasonic flowmeter capable of measuring instantaneous blood velocity profiles and hence volume flow is illustrated in Fig. 4 (6). In this system, the exciter provides a short (1.0 microsecond) burst of high-frequency (6.0 megahertz) excitation to the single piezoelectric transducer. For a relatively long interval (50 microseconds) after each transmit pulse, the scattered and Doppler-shifted signals returning to the transducer as the ultrasonic pulse traverses the vessel are processed by the receiver circuit and telemetered to the external electronics. Because the velocity of sound in blood is well known, the Doppler frequency shift in the returning signals at any instant corresponds to blood velocity at a particular point in the lumen. Consequently the external electronics can provide an instantaneous blood velocity profile whose integral over the lumen diameter can give an accurate estimate of volume flow.

An additional feature of prime importance in this system is the integrated circuit command receiver which, under control of the external command transmitter, serves to disconnect the battery power supply from the flowmeter electronics per se during periods when data are not required. In this manner, the useful operating life of the implantable unit can be enormously prolonged since the power drain of the command receiver is approximately 0.1 percent of that of the flowmeter.

For blood pressure telemetry, a thin silicon diaphragm approximately 1.0 millimeter in diameter and 5 micrometers thick is used (7). Tiny silicon resistors formed in this diaphragm using integrated circuit fabrication processes are sensitive in value to the blood pressure deflecting the diaphragm. One or more of these silicon piezo-resistive pressure transducers can be implanted along with signal processing and telemetry electronics similar to the electronics used in the Doppler flowmeter, to monitor blood pressure at various sites in the cardiovascular system. Such implantable telemetry systems offer considerable promise in medical research involving animal models of human disease.

Prostheses

For many humans the quality of life is greatly diminished by loss of some natural function. Blindness, deafness, paralysis, and loss of limb are afflictions suffered by many. The powerful and compact sensory, computational, and display capabilities of integrated electronics make possible exciting new avenues for prostheses to remedy these functional deficiencies. Some prostheses, especially sensory prostheses, operate transmodally, mapping the information normally gathered by the impaired sense onto an unimpaired sense. The optical-to-tactile reading aid described below is such a prosthesis. Others, such as the cardiac pacemaker, inject synthetic signals into the nervous system to replace missing natural signals. We now present examples of modern prostheses, possible only through integrated electronics, which are harbingers of things to come.

A transmodal sensory prosthesis. In principle, the Optacon is a direct translation reading aid for the blind which converts an optical image of a printed character on a page of an ordinary book, magazine, or newspaper to a vibrating tactile facsimile (*8*). By means of a simple optical system contained in a hand-held camera, an image of the printed character is focused on an integrated circuit array of phototransistors. Output signals from the phototransistor image sensor array are processed in a simple electronic system, and then used to control a corresponding array of piezoelectric tactile stimulators or bimorphs. Tiny pins cemented to the tips of the bimorphs protrude slightly through perforations in a rectangular plastic plate. A blind reader whose fingertip is resting over the perforations can then feel an exact tactile facsimile or vibrating image of the original printed character.

A photograph of the Optacon is shown in Fig. 5. The tiny camera in the right hand of the reader is scanned freehand across a line of print. Batteries and electronics are contained in the portable carrying case. The left index finger of the reader is covering the tactile display. Custom integrated electronics is crucial in two features of this instrument. The first is the 6 by 24 phototransistor image sensor array for the camera, and the second is the high-voltage driver circuits that are necessary to excite the piezoelectric transducers of the tactile display. Several thousand Optacons are now in daily use by blind individuals in many countries.

Neural substitutes. The signals that travel the nervous system are electrical in character, so it should in principle be possible to replace defective nervous tissue with an electronic substitute. The two major technological difficulties are (i) interfacing with the nervous tissue so that signals may be inserted or extracted and (ii) achieving in acceptably small volume and at acceptably low power the signal processing function of the tissue to be replaced. The impact of integrated electronics on both these obstacles has been revolutionary.

That electronic substitution for defective neural tissue is in practice feasible is proved by the success of the cardiac pacemaker. Approximately 100 cardiac pacemakers are installed daily in the United States, and their manufacture and sale is a $100-million business. The cardiac pacemaker, however, is required to achieve only a relatively simple function, that of assuring the arrival of a periodic stimulation at a single extended site. It was possible to achieve this

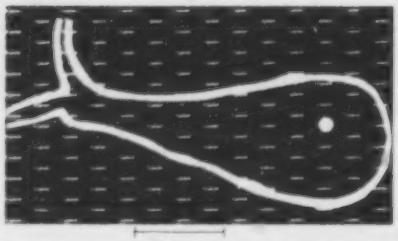


Fig. 6. Photomicrograph of semiconductor memory chip with image of cerebral neuron superimposed. The bar represents 50 μm.

simple function in an acceptable volume and at an acceptable power level even before the advent of integrated circuits. So also was it possible to achieve a pacemaker for the lung (*9*) for those paralyzed by a spinal cord injury or suffering from sleep apnea, and to develop a peroneal nerve stimulator (*10*) to remedy foot drop and improve the gait of hemiplegics. In all these cases a simplified version of the neural function replaced suffices to achieve an important or vital objective. One may now ask, however, to what degree our horizons in neural prosthesis have been extended by the advent of integrated electronic

Figure 6 is a photomicrograph of a commercially available integrated circuit memory chip; superimposed is the image of a cerebral neuron. Dozens of active semiconductor elements fit within the area of the cell body. The two-dimensional density of semiconductor memory or logic elements of advanced but current design is more than an order of magnitude higher than the corresponding density of cerebral nervous tissue. Even allowing for less efficient packing in the third dimension, it is now possible to attain a volume density of integrated circuit logic elements which is comparable to that of natural nervous tissue and to execute logical functions of similar complexity at similar energy cost. It is now technologically feasible to replace defective nervous tissue with synthetic neural

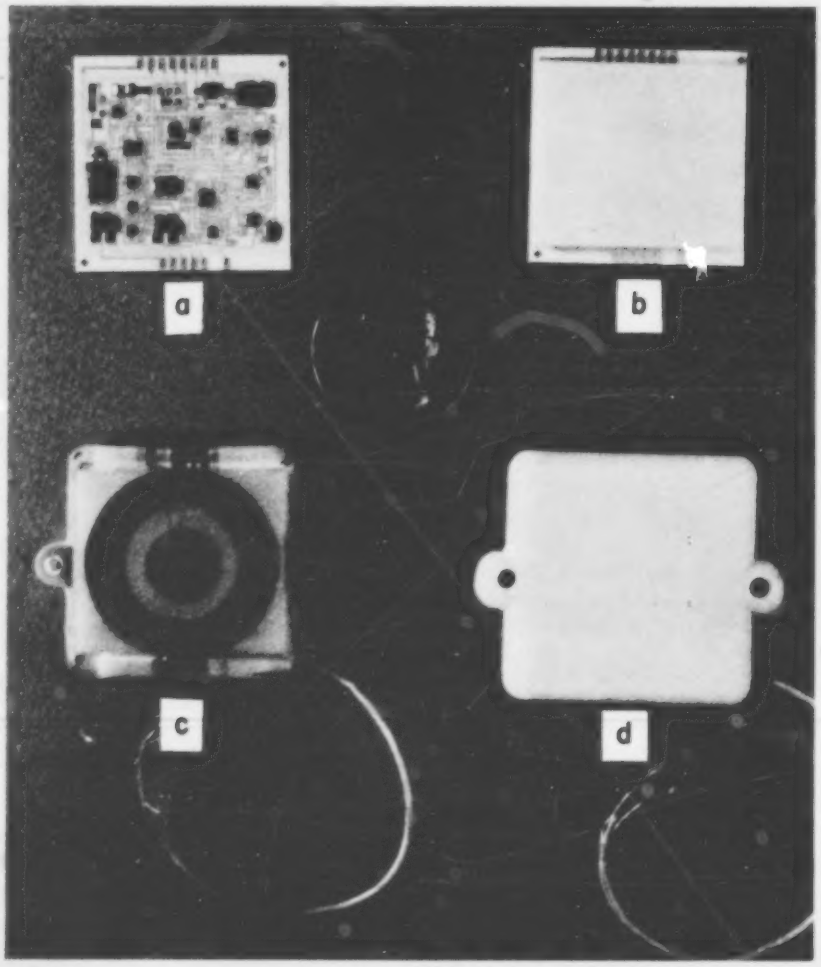


Fig. 7. The STIMULISS IV B auditory prosthesis implantable receiver-stimulator unit. (a) Hybrid microcircuits, (b) hermetically sealed, (c) with radio-frequency and ultrasonic receiver antennas, and (d) fully encapsulated.

circuits of similar size and power needs; the implied possibilities for prostheses are countless.

The most striking of the neural prosthesis possibilities opened by the high intelligence density of integrated circuits are sensory prostheses for hearing and vision. In both cases, the information flow rate is high, and the stimulation pattern which must be generated is complex in both space and time. One of the first cases to which integrated circuit technology has been seriously applied is that of an auditory prosthesis, an implantable inner ear for the profoundly deaf.

Basically the ear is a transducer that converts the mechanical energy of a sound wave (in which information is encoded) into a patterned array of electrical impulses traveling up the auditory nerve. In a significant number of cases of profound deafness the transduction machinery of the cochlea is defective, but the auditory nerve itself remains viable. It has been known for some time that electrical stimulation of the auditory nerve produces the sensation of sound in such cases. Considerations based both on auditory physiology and on speech theory indicate that, given the restricted bandpass characteristic of a single neural channel, the independent stimulation of eight to ten subsets of nerve fibers of the auditory nerve is required to convey the information content of speech. An eight- to ten-electrode array of stimulation microelectrodes must therefore be inserted into the auditory nerve. Because of the danger of infection, the microelectrode array cannot be driven on a permanent basis by wires running through the skin, but must be driven by a fully implantable receiver-stimulator unit. A series of cochlear prostheses are under development to achieve these functions (*11*). Integrated electronics technology is crucial to the fabrication of a small low-power implantable receiver-stimulator microelectrode array and a small low-power supercutaneous speech processor-transmitter. A prototype four-channel implantable unit has been developed and tested in vitro and in animals, and the fabrication of microelectrode arrays using photolithographic techniques as well as the design of the speech processing algorithm is in progress. Figure 7 shows the first-generation four-channel unit in several stages of construction. Power is transmitted on a radio-frequency link; the receiver coil is apparent. Data are transmitted in digital form over an ultrasonic link for noise and misalignment immunity; the receiver crystal is also visible (in Fig. 7) within the radio-frequency coil.

The development of a visual prosthesis for the blind offers even greater technological challenge, since the data acquisition rate is two orders of magnitude greater than that for the ear and there are the dimensions of color, flicker, and relative position of the eye to the head to be dealt with (*12*). The greatest obstacles to the visual prosthesis at present are, however, probably not technological but physiologic. One major obstacle is that the optic nerve is not viable in most cases of blindness (in contrast to the auditory case where nerve survival is common). Stimulation of the surface of the visual cortex produces bright spots (phosphenes) in the visual field, and most of the efforts on visual prostheses have been based on this phenomenon. There are, however, two basic problems confronting the cortical visual prosthesis. First, that part of the cortical surface corresponding to the central high-acuity portion of the retina is buried deep in a cortical fold, making surgical access all but impossible. Second, the cortex is the site of the final stages of optical signal processing and is not in principle a good place to insert what is essentially input data on the visual field. It is possible that the required electronic function for a visual prosthesis can be accomplished within the appropriate space and power constraints if the physiologic and anatomical problems of appropriate insertion point can be solved.

Sensory prostheses have been singled out as being complex and requiring the capabilities that integrated circuits offer. In addition, sophisticated skeletal muscle prostheses could require comparably complex signal processing (*10*). Consider, for instance, a prosthesis which might operate a paralyzed arm through neural stimulation. An array of muscles would have to be stimulated in a programmed fashion, and in a manner influenced by sensory feedback on limb position, grasp pressure, and the like. Some progress has been made in this direction, but only with the development of integrated circuits can the technological problems involved be surmounted.

Summary

There are two salient points which convey the essence of the impact of integrated electronics in medicine. First, the quality and availability of health care are becoming increasingly dependent on radically new diagnostic, monitoring, and prosthetic instruments provided by electronics. Second, many of these new instruments, such as ultrasonic imaging systems, blood flowmeters, and artificial inner ears, impose very stringent performance requirements, which can be met only through innovative application of integrated electronics. In summary, the initial impact of integrated electronics on health care is now visible, and consequently it is apparent that we are beginning a new era of revolutionary advances in medical instrumentation. Health care may indeed present the most promising opportunity to improve the quality of life in our society through electronics.

References and Notes

1. EMI, Electrical and Musical Instruments Ltd., Hayes, Middlesex, England.
2. O. Z. Roy, *Crit. Rev. Bioeng.* **2**, 259 (1975).
3. J. R. Cox, Jr., F. M. Nolle, R. M. Arthur, *Proc. IEEE* **60**(10), 1137 (1972).
4. J. Shott, R. Melen, J. Meindl, *Digest of Technical Papers, 1976 International Solid-State Circuits Conference*, p. 200.
5. C. F. Hottinger and J. D. Meindl, *Proc. IEEE* **63**(6), 984 (1975).
6. R. W. Gill and J. D. Meindl, *IEEE J. Solid-State Circuits* **SC-10**(6), 464 (1975).
7. Samaun, K. D. Wise, J. B. Angell, *IEEE Trans. Biomed. Eng.* **BME-20**(2), 109 (1973).
8. J. G. Linvill and J. C. Bliss, *Proc. IEEE* **54**, 40 (1966).
9. W. W. Glenn *et al.*, *J. Thorac. Cardiovasc. Surg.* **66**, 505 (1973).
10. D. R. McNeal and J. G. Resnick, *Adv. Biomed. Eng.* **6**, 209 (1976).
11. T. R. Gheewala, R. D. Melen, R. L. White, *IEEE J. Solid-State Circuits* **SC-10**(6), 472 (1975).
12. T. D. Sterling, Ed., *Visual Prosthesis* (Academic Press, New York, 1971).

13. The support of NIH grants GM-17940 and NS 10532-05 and contract N01-CG-3-2005 in the preparation of this article is acknowledged.

Medical Electronics (I): Computerized X-ray Scanners

Arthur L. Robinson

Medical science tends to advance incrementally, and full-fledged breakthroughs are rare. The discovery of the x-ray by Wilhelm Conrad Roentgen in 1895 and the subsequent development of the science of radiography is one notable example. Since 1972, a new x-ray device known as the CAT-scanner (for computerized axial tomography) has been appearing in an increasing number of hospitals and clinics. On the basis of their experience so far, many radiologists are saying that these computerized x-ray scanners are the greatest advance in diagnostic medicine since Roentgen's discovery, while others are only somewhat less effusive in their praise. CAT-scanners have had an indisputably marked effect on the way radiologists and surgeons diagnose their patients, but it is still too soon to evaluate what the overall contribution of the scanners to the quality of health care will be.

The enthusiasm for CAT-scanners derives from their superior ability to detect abnormalities (lesions) in the brain as compared with such conventional neuroradiological techniques as standard skull x-radiography (roentgenography), angiography, pneumoencephalography, and radionuclide scanning. Radiologists also cite the relatively noninvasive character of the scanners and their potential for reducing the cost of health care for patients who otherwise would be hospitalized.

In the diagnosis of numerous abnormalities of the brain, radiologists at the Mayo Clinic have reported an overall error rate with CAT-scanning of 4 percent on 12,000 scans over a little more than 2 years, for example (*1*). Disorders visualized included brain atrophy, degeneration of the brain, hydrocephalus, cysts, tumors of the brain and the eye, infarcts (dead areas of the brain due to loss of blood supply), and hemorrhage (Fig. 1). In addition, they find that CAT-scanning is applicable to all of the above-mentioned categories of abnormalities, whereas the other methods are each limited to certain ones only.

In conventional x-radiography, the image obtained on a film after a diverging x-ray beam passes through the subject is a projection or shadow of everything standing between the x-ray source and the film. Thus, the image may contain many overlapping organs and tissues which are difficult to separate. In addition, whereas an observer can easily distinguish between air, soft tissue, and bone in an x-ray photograph, the same viewer cannot easily see the few percent difference in the attenuation of x-rays by normal and diseased tissue, even when overlapping images are not a complicating factor.

The method embodied in computerized x-ray scanners to overcome these difficulties is a specific example of a general mathematical technique called reconstruction of images from projections. In principle, if x-ray photographs are made of a person's head at an infinite number of angles, it is mathematically possible to reconstruct a full three dimensional image of the skull and its contents from these projections. Such reconstructions can be made from a finite number of projections, but the reconstructed image is no longer exact.

A number of researchers have made reconstructions of two-dimensional cross sections normal to an axis of rotation of an object (transverse axial tomography) from x-ray photographs taken at equal angular intervals around the axis. This procedure overcomes the problem of overlapping, but the cumulative x-ray dose to a patient would be excessive. In addition, scattering of x-rays by parts of the patient's body would cause a loss of contrast, as it does in conventional x-ray radiography. The use of an electronic detector in place of the x-ray film together with a collimated, narrow x-ray beam and computer processing solves these problems.

Fig. 1. Reconstruction of the head of a patient with a calcified cyst. The view is from the top of the head down toward the body, and the cross section is of the mid-skull. Easily seen are the lateral ventricles, quadrigeminal cistern, and the cyst (white area in the lower right quadrant). The white area outlining the head is the skull. [Source: Ohio-Nuclear, Inc.]

Since the detector records only a small region at a time, in order to duplicate the area recorded in an x-ray photograph, the x-ray source and the detector must scan the region to be imaged. In the first generation scanners, the x-ray source and the detector scan together normal to the axis of rotation of the object, and thus generate a series of parallel x-ray beams in the plane of the cross section to be reconstructed. Only x-ray photons not scattered out of the beam are detected. Readings of the attenuated x-ray beam during the scan are stored in a minicomputer. At the end of a scan, the frame that holds the x-ray source and the detector rotates 1 degree, and another scan begins.

The computer completes the reconstructed image of the cross section either after or as the data from 180 or more scans accumulate, depending upon which of several possible algorithms it uses (*2*). The image consists of a rectangular array of elements, each of which represents an area of the cross section about 1.5 millimeters on a side. A cathode-ray tube or television screen displays the image. The cross section is not mathematically thin; its thickness is determined by the thickness of the x-ray beam and is 8 or 13 millimeters.

In the simplest algorithm, the brightness of each element represents the sum of the total attenuations of each x-ray beam that passes through the element. The method used by scanner manufacturers involves modifying the projections, so that this line summation gives a closer approximation to the true attenuation in each element.

With computerized x-ray scanners, differences of absorption as small as 0.5 percent can be distinguished, because the entire range of attenuations need not be displayed simultaneously, as on film (Fig. 2). By selecting a small range of attenuations to be displayed in the reconstruction, the viewer can easily pick out small changes

The author is a member of the staff of the Research News Section of *Science*.

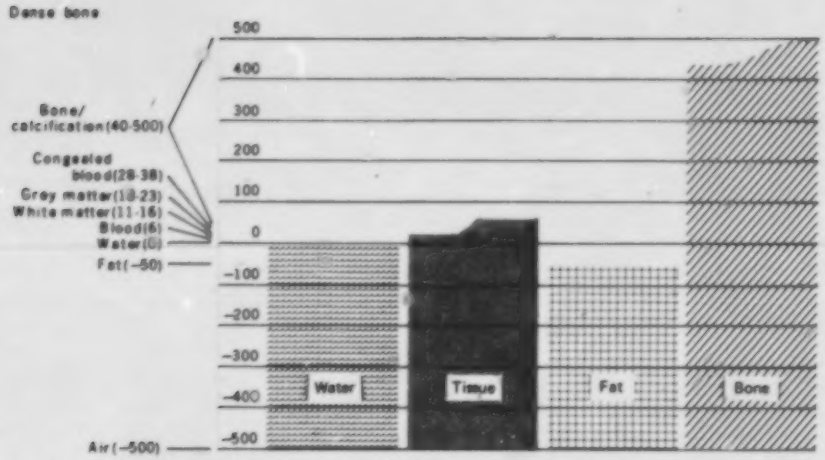


Fig. 2. Absorption coefficients for x-rays of various anatomical materials on an arbitrary scale of 1000. With computerized x-ray scanners, only a limited portion of this range need be displayed in the reconstructed image, thus greatly increasing the "density resolution" of reconstructed cross sections as compared with conventional x-rays photographs, which must display the entire range at once. [Source: EMI Ltd.]

that would be missed in a normal x-ray photograph. The spatial resolution of the image, however, is not as good as that of an x-ray photograph, being limited by the size of the picture elements. The accumulated radiation dose from a CAT-scanner is comparable to that from a series of skull x-rays. A set of three or four x-rays imparts a dose of 2 to 4 rads, as does a series of three or four sets of scans that constitutes a CAT examination. (In actual machines, two detectors are used, so two cross sections are obtained simultaneously. Thus, a typical examination results in six or eight cross sections.)

A careful evaluation of the efficacy of CAT-scanners in detecting and differentiating lesions as compared with other diagnostic techniques has yet to be reported. Many hospitals and clinics, however, have accumulated what statisticians call anecdotal data on the performance of the scanners. According to Hillier Baker at Mayo, for example, the effect of computerized x-ray scanners on the use of the other techniques has been substantial. Pneumoencephalography has decreased in use by 90 percent, angiography by 10 to 15 percent, and radionuclide scans by 60 percent.

Figures such as these cannot be taken as typical because the frequency of use of each method depends on the type of patient a clinical center treats. Thus, while pneumoencephalography has been the most drastically effected diagnostic technique in most (but not all) centers, a more typical statement of the impact on angiography and radionuclide studies is that the CAT-scanner can be used as a screening tool to select patients for whom the other methods may then be necessary.

Angiography involves the injection of a radiopaque medium into the arterial blood system, so that the cerebral vascular system is made visible to conventional x-rays. The substances used in angiography, however, are irritating to the body, and the probability of complications or even death is uncomfortably high (about 1 percent). Even so, radiologists continue to do angiography, because of its superiority in visualizing some kinds of malformations of the cerebral vascular system.

Pneumoencephalography is an especially unpleasant procedure to undergo. The cerebrospinal fluid is removed and replaced with air or another gas. Air in the spaces normally occupied by the fluid, such as the ventricles of the brain, acts as a high contrast medium in the sense that the x-ray attenuation is greatly reduced. The after-effects are severe and include headache, nausea, and vomiting. Both angiography and pneumoencephalography may impart high radiation doses, from 25 to 50 rads, and both require hospitalization.

Fig. 3. Reconstruction of the abdomen of a normal patient. The view is up through the body from the feet toward the head. Some organs visualized are: *B*, bowel; *P*, pancreas; *L*, liver; *A*, aorta; *S*, spleen, and *SC*, spinal column and canal. Surrounding the abdominal cavity are the skin (white ring), fat (dark ring), and muscle (especially prominent behind the spinal column). [Source: EMI Ltd.]

In radionuclide scanning, a gamma-ray emitting isotope is injected intravenously. If a lesion, such as a tumor, is associated with a breakdown of the blood-brain barrier that separates the blood from the tissue of the nervous system, the isotope can leak out and accumulate in the area of the lesion. The radiologist can obtain a map of gamma-ray intensity with position and thus locate the lesion. The dose of radiation received is quite low, the procedure is one with a relatively low risk, and hospitalization is not required, but the spatial resolution and the sensitivity for certain lesions is low.

In CAT-scanning, the patient lies on a motorized couch with the head between the x-ray source and the detectors. In the most used version of the scanner, the patient's head fits into a flexible rubber bag which projects into a water-filled box, so that all x-rays enter and exit through the water. The water ensures that even x-rays not passing through the head will be attenuated to some extent and thus acts as a reference medium. The water is also said to reduce the requirements on the performance of the x-ray detectors and to reduce the number of artifacts in the reconstruction.

The procedure is relatively comfortable for the patient, but the subject must remain motionless for the 5 minutes it takes to complete the scanning for each set of two cross sections in order to avoid so-called motion artifacts in the reconstructed image. This restriction can pose problems for very sick or uncooperative patients. Many radiologists now use a contrast medium to enhance the image of lesions, especially when tumors are suspected. The contrast medium is an iodine-containing compound which accumulates at certain types of lesions in a manner similar to radioisotopes. Severe allergic reactions can occur, but the incidence of mortality is less than one in twenty thousand, for example; thus this partially invasive procedure is considered quite safe. CAT-scans can be made on an outpatient basis, so the cost of a several days' stay in a hospital are not incurred. Outpatient savings averaged $1490 per patient according to a recent study car-

ried out at the University of Toronto.

Radiologists report that CAT-scanners have certain limitations. Metal clips, for example, from previous operations cause artifacts in the reconstructed image. The most serious limitation seems to be the inability to detect small lesions that are close to bone, a situation common in the region at the base of the skull. Called a partial volume artifact, the failure to image these lesions stems from the relative thickness of the cross section and from the large difference between the x-ray attenuation of bone and tissue. If a bone occupies part of the volume of a picture element, its high density obscures the small difference between the attenuation of a small lesion and normal brain tissue in the picture element.

Researchers in widely separated fields have rediscovered methods for image reconstruction many times. Radioastronomers, for example, use image reconstruction methods when they construct high resolution maps of radio sources by aperture synthesis from data obtained with arrays of antennas (*Science*, 13 June 1975, p. 1071). Electron microscopists have been responsible for much of the recent mathematical development of image reconstruction through their research on the structure of biological specimens with the use of Fourier microscopy. And in 1917 the Austrian mathematician J. Radon laid the mathematical basis for the algorithms used by the scanner manufacturers.

Although the basic ideas of CAT-scanning had been developed by the early 1960's, it was not until 1972 that a commercial scanner was available, perhaps because of a lack of small and inexpensive computers. Godfrey N. Hounsfield of the Central Research Laboratories of EMI Ltd., Hayes, Middlesex, England, headed the development of an apparatus designed for computerized x-ray scanning of the head. This device, called the EMI-Scanner, has been largely responsible for the enthusiasm for CAT-scanners and has inspired the emergence of several competitor scanners. According to the Bureau of Radiological Health, Rockville, Maryland, four companies have filed the radiation safety documents required for marketing instruments in the United States.* (Since this article first appeared, several more CAT-scanners have appeared. A recent meeting, for example, featured reports on the imaging capabilities of 14 different machines. Most of these machines can image both the head and body and have scanning times as short as 5 seconds.)

The potential market for CAT-scanners is very large. Observers say that hospitals with 200 or more beds provide the most likely buyers of such devices for now. Some optimists, however, are speculating that within a decade, a CAT-scanner will appear nearly everywhere that conventional x-ray machines exist now. The total number of scanners sold is approaching 500, a little more than 200 of which are in operation. (As of December 1976, the number of scanners installed world-wide had climbed to 600, about half of which were in the United States.)

The primary advantage of many of these new machines will be speed, so that chest motion due to respiration will not degrade reconstructed images of the torso and so that an increased number of patients can be accommodated. For example, EMI now has in operation three prototypes of a body scanner that can accumulate the data for one cross-section reconstruction in 20 seconds, the length of time a healthy person can easily hold his breath.

The prospect of imaging cross sections of any part of the body is highly exciting (Fig. 3). Radiologists seem to agree, however, that the clinical efficacy of the available body scanners is still unproved. So far, these scanners have been used primarily in a research environment where their capability to detect abnormalities already located by more conventional means is being ascertained. (This situation has not changed greatly. Recently, for example, radiologists at the Cleveland Clinic reported that CAT-scanners seemed to more accurately spot tumors of the pancreas, a notoriously difficult organ to image, than either ultrasonic techniques or angiography, but adequate statistics to be sure of this conclusion had not yet been accumulated.)

If the value of body scanners is yet to be shown, the demand for computerized x-ray scanners for neuroradiology is overwhelming. It is not uncommon for hospitals having CAT-scanners to operate them up to 16 hours a day, 5 or 6 days per week. For a basic study of three or four head scans, two patients per hour can be accommodated. Even so, waiting time to receive a CAT-scan can be as long as 4 weeks, and scheduling of emergencies is often a problem. The price for a CAT-scan varies from hospital to hospital and depends on such variables as whether contrast studies are done, but the average price for an examination is between $200 and $250.

The present scanners cost an average of $385,000 (plus a $25,000-per-year service contract to keep them operating), and the new fast body scanners may cost $550,000 or more. Because of this cost, observers say that as the scanners proliferate and as more and more normal patients (those in which no abnormalities are detected) are scanned, the cost of health care will rise, as reflected, for example, in the cost of health insurance. These observers question whether this cost will translate into a quantifiable improvement in health care, despite the acknowledged diagnostic capabilities of the brain scanners. To those who have to wrestle with health care planning and who do not see the patient on an individual basis, the issue may be a matter of balancing priorities. But to those who have suffered through pneumoencephalography, the value of the CAT-scanner may be more obvious.

*In addition to EMI, the companies are Digital Information Systems Company, Silver Spring, Maryland (now called Pfizer Medical Systems, Inc.), Ohio-Nuclear, Inc., Solon, Ohio, and Neuroscan, Inc., Newport Beach, California. The Pfizer (ACTA) scanner and the Ohio-Nuclear (Delta) scanner can image the whole body, as well as the head.

References

1. H. L. Baker, Jr., O. W. Hauser, J. K. Campbell, D. F. Reese, C. B. Holmen, *J. Am. Med. Assoc.* **233**, 1304 (1975).
2. R. Gordon and G. T. Herman, *Int. Rev. Cytol.* **38**, 111 (1974).

Medical Electronics (II): Computerized Scanner Explosion

Arthur L. Robinson

Computerized x-ray scanners have had a dramatic impact on the practice of neuroradiology since their introduction 3 years ago. Efforts are now under way to extend the range of applicability of image reconstruction from projections (the principle on which x-ray scanners are based) in medicine still further. In particular, researchers are working on fast x-ray scanners, on the combination of image reconstruction with nuclear medicine, and on low dose methods of imaging.

As now performed, the scanning procedure is necessarily slow. The basic scanning action consists of the motion of a narrow x-ray beam and a single detector across the head or body of a patient; this linear scan is followed by a rotation of the frame holding the x-ray source and detector about the patient and another linear scanning motion. A two-dimensional image of the cross section of the head or body in the plane that is traced out by the x-ray beam during the scanning is typically reconstructed from 180 or more one-dimensional projections, one of which is produced by each linear scan (*Science*, 7 November, p. 542).

The time to read the detector 40,000 or more times and the time to move the x-ray source and detector repeatedly through the linear and rotational motions together add up to about 4.5 minutes. Ways to reduce this time include the use of many detectors in the plane of the cross section, so that many data points are taken simultaneously, and the reduction or elimination of all mechanical motion.

One reason for needing faster scanners is that any motion of the patient during the scanning introduces artifacts into the reconstructed image, and it is often difficult for sick patients to remain sufficiently immobile for 5 minutes. Even chest motion during respiration could be overcome with 5- to 20-second scanning times because the breath could be held that long. And more patients could be accommodated if the examination time were reduced.

Manufacturers and academic investigators who are now in the process of building computerized x-ray scanners that are equipped with multiple detector arrays are also using an x-ray source geometry called a fan beam. In conventional x-radiography, the x-ray tube emits a diverging cone of radiation that may be 40 centimeters in diameter by the time it reaches the film. In the first generation of computerized x-ray scanners, the x-rays are collimated into a thin beam which, in conjunction with a small collimated radiation detector, minimizes the effect of scattered radiation, since x-ray photons scattered out of the beam by parts of the patient's body are not detected. Scattered radiation reduces the contrast obtainable in conventional x-radiography.

The fan beam is a kind of compromise in that the x-ray beam remains confined to the plane of the cross section being imaged, but it is allowed to diverge within the plane, thus tracing out a fan-shaped area and simultaneously striking some or all of the detectors in a one-dimensional array. Because the area of the detector array is much smaller than that of an x-ray film, the scattering problem, although not absent, is reduced manyfold. Groups at Stanford University working under Douglas Boyd and Albert Macovski, for example, have shown that such scattering can be negligible if additional collimation is added to the detector array or if the projection data are modified prior to the reconstruction.

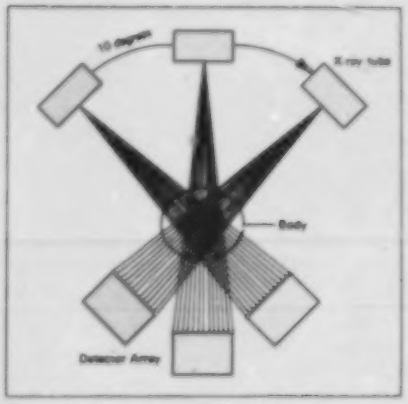


Fig. 1. Schematic diagram of fan beam geometry. In this example, rotational motion is in increments of 10 degrees. At each angular position, the x-ray tube and the detector tube scan linearly in the plane of the diagram to ensure that x-rays pass through every portion of the body. [Source: EMI Ltd]

Whole Body Scans in 5 Seconds

The number of detectors in scanners is increasing rapidly. Different models now available which have 1, 3, or 30 detectors have corresponding scanning times of 4.5 minutes, 2 minutes, and 20 seconds. These instruments retain a combination of linear and rotational scanning motions, although the rotation increments have increased from 1 to 3 to 10 degrees, respectively (Fig 1).

Even faster x-ray scanners will retain only the rotation motion. One brain scanner just becoming available has 128 detectors and rotates continuously over 360 degrees in 9 seconds. A number of researchers are working on whole body scanners that will be able to collect all the data needed for the reconstruction of a cross section in 5 seconds with the aid of 250 or more detectors.

Many of these advanced computerized x-ray scanners collect more data than the earliest versions so that larger areas of the body can be imaged with higher resolution. As a result of the large amount of data processing, the time a computer algorithm requires for the reconstruction does not match the data acquisition time, varying from 2 to 5 minutes per reconstruction. In the future, the use of faster algorithms, multiple computer processors (one to supervise data collection and one to carry out the reconstruction algorithm), and special-purpose, hard-wired (nonprogrammable) processors should reduce this time.

Some of the fast scanners will also use a variety of detector that potentially could make the computerized scanner less expensive (an average of $385,000 for the slower scanners and upward of $500,000 for the first fast whole body scanner). The early scanners have had detectors consisting of a scintillating crystal, such as sodium iodide, and a photomultiplier tube. The cost and complexity of an array of these detectors can rapidly increase when very many are used. One alternative is a multiwire gas-filled chamber. Operation of the multiwire chamber is the same as that of a conventional single-wire gas chamber, but an array of wires makes it possible to determine the location of the x-ray photon being detected. Boyd's group at Stanford has built a multiwire proportional chamber filled with xenon

The author is a member of the staff of the Research News Section of *Science*.

gas at high pressure, for example, and David Chesler and his colleagues at Massachusetts General Hospital in Boston, who are building a 5-second whole body scanner, are experimenting with a similar detector. Some commercial scanners also will have gas detectors.

Ultimately, machines having arrays of many x-ray sources and many detectors will avoid all mechanical motion whatsoever. A start in this direction has been made by researchers in the Biophysical Sciences Laboratory, headed by Earl Wood, at the Mayo Foundation in their efforts to make cross-sectional images of the beating heart. According to Richard Robb at Mayo, motion pictures of multiple cross sections of a beating heart of a dog have been made by electronically synchronizing the heart beat with x-ray pulses. The investigators use a fluorescent screen, an image intensifier, and a television camera to record two-dimensional projections at the rate of 60 per second. From the projections obtained at 36 viewing angles, cross-sectional reconstructions at many anatomic levels of the heart can be made for each instant of time and displayed to show the dynamic changes in the size and shape of the heart.

The investigators' next step will be to use perhaps 30 x-ray sources along with image intensifier-television detectors positioned in a semicircle. By making 30 x-ray images simultaneously at intervals of 1/60 second, electronic synchronization would not be required, and the technique could be applied to visualizing the human heart without the effects of heart motion obscuring the image.

Sometimes it would be useful to obtain images of body or head sections other than the transverse axial cross sections obtained from x-ray scanners. In principle, other orientations could be derived from a set of cross sections, but the thickness of the cross sections (up to 13 millimeters) would limit the usefulness of such images. William Glenn at Massachusetts General Hospital and Samuel Dwyer, III, at the University of Missouri, Columbia, have collaborated on a way to overcome this problem by making several overlapping cross sections with a conventional x-ray scanner. From 16 reconstructed cross sections each 8 millimeters thick and covering a total thickness of about 3.5 centimeters, a computer can create a full three-dimensional image with a resolution of about 2 millimeters and display two-dimensional sections at several orientations. Time varying displays can also be made so that a rotating image of the ventricular system of the brain can be shown, for example.

The image obtained in radionuclide scanning is analogous to an x-ray in that it is a two-dimensional projection. It is unlike an x-ray in that the source of radiation is a gamma-ray emitting isotope localized within the patient's body—that is, it is an emission rather than a transmission technique. By applying methods similar to those used with computerized x-ray scanners, however, image reconstruction from projections can be combined with nuclear medicine. In fact, David Kuhl and his coworkers at the University of Pennsylvania were using image reconstruction methods before the x-ray scanners were developed, and their instrument has evolved through several stages.

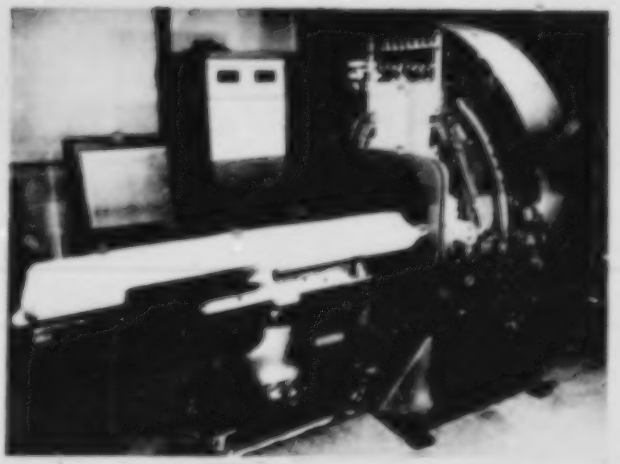


Fig. 2. The whole body positron emission transaxial tomograph designed at Washington University, St. Louis. The patient lies on the couch, which slides into the detector array with eight detectors in each side of the hexagon. The scanning motion consists of a 5-centimeter linear scan followed by a 3-degree rotation. This sequence is repeated until 60 degrees are covered. [Source: Michael Phelps, Washington University]

The present scanner has 32 coplanar scintillation detectors in a boxlike arrangement that surrounds the patient's head. The detector array rotates continuously, and one reconstructed cross section can be obtained in 50 seconds up to a few minutes. The resolution of the reconstructed image is about 1.6 centimeters and, as in most nuclear medicine techniques, is much coarser than the 1.5-millimeter resolution of x-ray reconstructions. According to Kuhl, the reconstructed image correctly assays the radioactivity to within 5 percent, which is more than adequate, since the objects being imaged have a high contrast. (The computerized x-ray brain scanners advertise an ability to see differences in x-ray attenuation of 0.5 percent, but the difference in attenuation between normal and diseased tissue may be only a few percent.)

Kuhl and his associates have used their scanner for diagnosing various cerebral lesions. In another study, they made three-dimensional maps of local cerebral blood volume by reconstructing cross-sectional maps of the intensity of gamma-ray emission from red blood cells labeled with technetium-99, which had been injected intravenously.

Other researchers have made cross-sectional reconstructions from gamma-ray emission with the use of a gamma (or Anger) camera. A gamma camera is a large (up to 38 centimeters in diameter) position-sensitive scintillation detector and is often used for conventional radionuclide scanning. Thomas Budinger, Grant Gullberg, and their associates at the University of California, Berkeley, and at the Lawrence Berkeley Laboratory, for example, have imaged the brain, the heart, the kidneys, and the liver by rotating patients in front of the gamma camera and recording data at 36 or 72 different angles. Since the detector is two-dimensional, the data necessary for as many as ten cross sections are obtained simultaneously. Such a system is not suitable for clinical use, however, because sick patients cannot always be moved, and because it is slow.

If attenuation of the gamma rays as they pass through the patient's body is severe, the image of the radioactive region may be distorted in position or in intensity. An imaging technique with the two gamma rays emitted in opposite directions with equal energies when an electron and a positron annihilate can overcome this difficulty. Sources of positrons are radioactive isotopes that are introduced into the patient just as in standard nuclear medicine procedures. In addition, the contrast of the reconstructed image is enhanced because scattered radiation is not detected. A technique called coincidence counting is used to ensure that a detection event is registered only when two gamma rays traveling in opposite directions are detected within a very short time by two opposing detectors. This limits the detected events to a well-defined region between two detectors placed on either side of the patient.

Michael Phelps, Michel Ter-Pogossian, and their associates at Washington

University, St. Louis, have developed an instrument they call a positron emission transaxial tomograph (PETT). Forty-eight sodium iodide detectors are arrayed around the patient in a hexagonal arrangement with eight detectors on each side of the hexagon (Fig. 2). Each detector is connected to the eight on the opposite side of the array for coincidence detection, a situation roughly analogous to the fan beam geometry of the fast x-ray scanners. Both linear and rotational scanning motions are required, and one reconstructed cross section with a resolution of from 1 to 1.5 centimeters can be completed in 2 to 5 minutes.

With the lastest version of their system, the Washington University researchers have imaged such organs as the heart, the brain, the liver, the pancreas, the spleen, and the kidneys of healthy subjects. They have also detected tumors, arterial-venous malformations, and strokes in brain studies, and have seen myocardial infarctions in the heart.

Phelps is especially enthusiastic about the use of carbon-11, nitrogen-13, and oxygen-15 as positron sources. These isotopes of atoms common in organic materials can be used to label such compounds without disrupting their chemical or physiological behavior. Thus, the metabolism of the compounds can be observed by image reconstruction. Phelps says that conventional labeling isotopes are, in effect, foreign to the body and may affect the way the compound is metabolized. The "organic" isotopes are short-lived, and a cyclotron is needed in the hospital to generate them. At present, cyclotrons of the type required may cost $500,000, in part because only a few are sold in a year. Positron sources that do not require cyclotrons and thus can be used in any clinic include gallium-68 and rubidium-82.

In other respects, imaging with heavy charged particles proceeds in the same ways as with x-rays. Budinger, Crowe, and their colleagues have reported on cross-sectional reconstructions of the brain. Instead of a linear scanning motion, the investigators used a parallel beam of helium ions that is equivalent in the many parallel beams produced by the x-ray scanners. The method is not yet ready for clinical use, however, because the patient must be rotated in front of the beam which is obtained from the Berkeley 184-inch cyclotron.

An imaging system developed by G. L. Brownell and his associates at Massachusetts General Hospital can be used both for image reconstruction and conventional radionuclide imaging. Two-dimensional arrays each containing 127 detectors (scintillating crystals and photomultiplier tubes) sit on opposite sides of the patient. When the two arrays are used as a coincidence detector, the picture corresponds to a normal isotope scan. When the arrays rotate through 180 degrees, however, and readings are taken at intervals of 5 to 7 degrees, the information required for several cross sections is accumulated. So far, the Massachusetts General group has made cross sections of artificial test objects (phantoms) and of the brain in human subjects.

Researchers working under Victor Perez-Mendez at Lawrence Berkeley Laboratory, and under Leon Kaufman and Chang Lim at the University of California, San Francisco, have constructed and tested a positron annihilation image reconstruction system that involves no scanning motion at all. The reconstructed image, however, is not a true transaxial cross section, but is more akin to an earlier form of tomography known as focal plane imaging. According to Kaufman, the entire system could cost much less than the least expensive x-ray scanner.

The multiwire proportional chambers used in the coincidence detector are a major factor in the cost estimate. The detector is similar to the xenon detector that is expected to be used in some fast x-ray scanners, but the array is two-dimensional, covering an area 48 by 48 centimeters. Coincidence events are detected by elements in the two arrays opposite one another on either side of the patient.

Still other improvements to computerized scanners are being explored. Observers consider imaging techniques that do not impart the 2- to 4-rad dose of x-rays but retain the ability of x-ray scanners to make images with a high fidelity to be especially important because 2 to 4 rads is still too high for routine screening of nominally healthy persons. One way to reduce dosage is to find improved algorithms so that accurate images can be reconstructed from fewer data points. Research on combining image reconstruction from projections with ultrasound (which is considered safe because it produces no ionizing radiation) and with nuclear magnetic resonance are also under study for this reason.

Budinger, Kenneth Crowe, and their associates at the Lawrence Berkeley Laboratory have been studying the utility of radiography using helium ions and other so-called heavy charged particles as another approach to low dosage imaging. The interactions suffered by such particles do not result in their being absorbed as x-rays are, but are more like collisions between hard spheres in which a small amount of energy is lost in each collision. Thus, in heavy charged particle imaging, it is the energy of the transmitted particles rather than their number that is measured.

Satellite Communications

Advances in electronic technology have contributed to growth in satellite communications.

Burton I. Edelson and Louis Pollack

I invite all nations to participate in a communications satellite system, in the interest of world peace and closer brotherhood among peoples of the world.—JOHN F. KENNEDY, 24 July 1961

With these words, President Kennedy, in the early months of his Administration, took a considered step to accelerate the pace of the nation's space program. His words, contained in a "Policy Statement on Communications Satellites" (*1*), set in motion a train of events which would not only help the United States to regain space leadership but would benefit many nations in improved technology and commerce and provide a new means for international cooperation. His statement recognized the potential value of earth-orbiting platforms to provide communications services, established government policy to coordinate activities and carry out research and development, called for implementation by the private sector, and invited other nations to join in the effort.

President Kennedy's bold and prescient statement was based largely on prediction and promise but very little on the results of development or demonstrated technology. Some 16 years earlier, in 1945, Arthur Clarke had first foreseen the use of space stations in geostationary orbit for communicating to points on earth (*2*). Then Pierce of Bell Laboratories carried out a system analysis in 1954 (*3*). And in 1957 the first Sputnik demonstrated the possibility of the use of satellites for many applications. But by 1961, at the time of the Kennedy statement, the only relevant communications demonstrations had been the short-lived Courier 1B satellite in 1000-kilometer orbit and the Echo 1 balloon in a 1600-km orbit, the former demonstrating active "store and forward" teletype transmission and the latter passive reflection of powerful microwave signals from one earth station to another. The effective demonstration of active microwave communications (Project Telstar and Project Relay) were still a year in the future. The first launch to a geostationary orbit (Project Syncom) would be made 2 years later (*4*).

One year after Kennedy's statement, the U.S. Congress passed the Communications Satellite Act of 1962 (*5*), which authorized the formation of a private corporation to carry out this country's interest in commercial satellite communications. As a consequence, the Communications Satellite Corporation (or COMSAT) was formed in 1963. Then in 1964, COMSAT joined together with communications entities of other nations to form an international body known as INTELSAT, the International Telecommunications Satellite Organization, to provide satellite communications on a global basis. In 1965 the world's first commercial satellite, INTELSAT I, known as "Early Bird," was launched and operational telecommunications service along a single path between the United States and Europe was inaugurated.

Operational Progress

Today the INTELSAT system spans the globe. Early Bird has been supplanted by eight larger, more powerful satellites operating with 150 earth stations. The system shown in Fig. 1 provides communications along 500 paths located in 80 countries. The system carries 8000 full-time simultaneous two-way telephone conversations, telex, data, and facsimile traffic, plus television and leased transponders.

The growth rate of the INTELSAT system during its first decade of operation (Fig. 2) has been large. Growth stemmed not only from improvement in the existing service over previous means such as radio and cable but from the offering of entirely new services such as global television and wide-band data and to areas not previously served at all by existing means.

The reliability of system operations has increased over the past years, equaling or surpassing that of other modes. Individual links in the system operate at a reliability of 99.99 percent or more, with the total system reliability exceeding 99.9 percent (*6*).

Technological Progress

In order for the operational capacity, quality, and reliability of the system to advance so rapidly, it was essential not only for the need to exist but for the technology to be available. The technologies necessary for satellite communications did indeed advance at a rapid rate.

In contrast to the case for many fields of aerospace and electronics system development in which 7 to 10 years frequently elapse from the development of a concept to the full deployment of an operational system, the progress in INTELSAT, which has seen the development of four generations of satellites in a decade, seems phenomenal. The Early Bird satellite, first of the series (Fig. 3), weighed only 38 kilograms. With limited power and bandwidth, its communications capacity could sustain about 240 two-way telephone circuits.

Great improvements were made in the three succeeding generations of satellites so that the INTELSAT IV series (*7*) currently in use represents an order-of-magnitude improvement in most operating parameters, for example, prime power and bandwidth, and thus in operating capacity. The most effective technique for increasing capacity through the first four generations of INTELSAT satellites came from increased effective radiated power from stabilized earth-pointing antennas. As shown in Fig. 4, the per circuit-year cost decreased by a factor of 30 in a decade.

The next generation of satellites, INTELSAT V, is now under contract. The first of this series, to be launched in 1979 (Fig. 5), will be the first body-stabilized spacecraft in the INTELSAT series (*8*). Its capacity will be greatly increased over that of previous generations, in part because of the increased primary power available and to a larger extent because of the increased communications band-

Dr. Edelson is director and Mr. Pollack is assistant director of COMSAT Laboratories, Communications Satellite Corporation, Clarksburg, Maryland 20734.

width utilized. The INTELSAT V satellite will use not only the 6/4-gigahertz frequency band (6 Ghz is the receive band and 4 Ghz is the transmit band) used in earlier INTELSAT satellites but will inaugurate use of an additional 500-megahertz bandwidth available in the 14/11-Ghz band.

The distinguishing features of the INTELSAT V satellite are shown in Fig. 5. The body of the satellite, stabilized along three axes in space, is used as a platform on which to mount the large sun-oriented solar array and the complex of earth-oriented antennas. The INTELSAT V satellite is expected to carry projected global traffic through the mid-1980's. In this period the Space Shuttle may be introduced for launching satellites of all types including some of the INTELSAT V series.

Earth stations in INTELSAT are owned and operated by the individual participating countries. Since the purpose of the satellite system is to interconnect terrestrial systems of the various countries, the number of earth stations per country is small. Some countries operate in two ocean areas, and some of those with very large traffic streams have established diversity paths. Moreover, some islands or geographically separate areas require their own earth stations. Thus, the United States, for example, has two East Coast stations, two West Coast stations, and three island stations (in Puerto Rico, Hawaii, and Guam).

The first earth stations to enter the INTELSAT system were those built originally to operate experimentally with Telstar and Relay. With these small satellites and with Early Bird, very powerful and thus expensive earth stations with large antennas were required. Through its first 10 years, INTELSAT standardized on the use of such large antennas, generally with apertures of 30 meters, coupled with supercooled parametric amplifiers, to yield a ratio of antenna gain to system noise temperature (G/T) of 41.7 decibels per degree Kelvin (Fig. 6). Improvements in earth stations through the decade included the introduction of very-wide-band system elements such as high-power (up to 12 kilowatt) traveling wave tube (TWT) amplifiers and low-noise parametric amplifiers (capable of operation to 55°K without cryogenics). Cassegrainian antenna systems have been introduced with shaped surface reflectors yielding high gain-to-aperture efficiency. Other changes have been made so that today all amplification and signal-processing circuits (except the high-power amplifier) are based on the use of solid-state devices.

As the power of the satellites and the number of earth stations in the system increased, it became clear that for many applications smaller earth stations would be both suitable and cost-effective. In 1976, therefore, INTELSAT adopted a second earth station standard, smaller, less expensive, and tending to new sys-

Fig. 1. The INTELSAT system. The present global satellite communications system consists of eight satellites in orbit, four operational plus four spare satellites connecting 150 earth stations located in 80 countries over 500 paths.

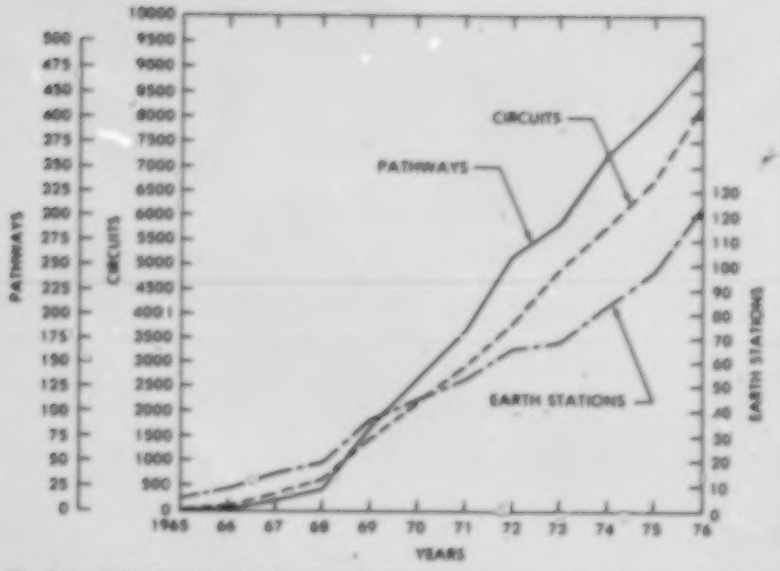


Fig. 2 (top left). Growth of the INTELSAT system. The number of earth stations, the communications paths connecting earth stations, and the traffic over these paths have each increased rapidly over the past decade. In many years each of these indexes of system growth increased by more than 50 percent. Fig. 3 (top right). The Early Bird satellite. Only one of these satellites was launched to provide initial transatlantic service. Early Bird proved the feasibility of the use of the geostationary orbit for commercial satellite communications. (Photo courtesy of Hughes Aircraft Company, Culver City, California) Fig. 4 (bottom). Technological progress in INTELSAT satellites. This table demonstrates the rapid technological progress in the series of INTELSAT satellites. Data are not shown for INTELSAT II, which saw service starting in 1967 with minor improvements over INTELSAT I, and INTELSAT IV-A, an improved version of INTELSAT IV, first launched in 1975 and currently in use. The INTELSAT I, II, IV, and IV-A satellites were built by Hughes Aircraft Company; INTELSAT III satellites were built by TRW Systems, Inc. The INTELSAT V satellite, now under construction for first launch in 1979, is being built by the Ford Aerospace and Communications Corporation. The INTELSAT satellites I through IV-A are spin-stabilized, whereas the INTELSAT V satellite will be body-stabilized along three axes.

tem applications. The new standard requires an aperture of about 11 m to provide a *G/T* of 31.7 *db/°K* with room-temperature parametric amplifiers. These stations are expected to be used widely in the INTELSAT system in areas with low-to-moderate traffic requirements. They will be particularly useful to inaugurate service to new areas, to provide service to second ocean regions, to furnish diversity routing, and also to provide domestic service.

Organizational and Economic Progress

Perhaps the most remarkable aspect of the global system is the institutional arrangements that make it possible. The INTELSAT system is an organization of 95 countries, all parties to an international agreement, equivalent to a treaty. The operating participants of INTELSAT are telecommunications entities in those countries—various government agencies, government-owned corporations, and private firms. The INTELSAT system carries on its activities through an international board of governors, a number of *ad hoc* and standing committees, a director-general with a full-time international staff, and a man-

	INTELSAT I	INTELSAT III	INTELSAT IV	INTELSAT V
YEAR OF FIRST LAUNCH	1965	1968	1971	1979
PRIME CONTRACTOR	HUGHES	TRW	HUGHES	FORD AEROSPACE
DIMENSIONS				
DIAMETER (m)	0.72	1.42	2.38	2.0
HEIGHT (m)	0.60	1.04	5.28	15.7
IN-ORBIT MASS (kg)	38	152	700	967
LAUNCH VEHICLE	THOR - DELTA		ATLAS - CENTAUR	
PRIMARY POWER (WATTS)	40	120	400	1200
TOTAL BANDWIDTH (Mhz)	50	500	500	2300
CAPACITY (TELEPHONE CIRCUITS)	240	1200	4000	12000
DESIGN LIFETIME (YRS)	1.5	5	7	7
COST/CIRCUIT YEAR ($)	22500	2000	1200	800

49

agement services contractor, a U.S. corporation (COMSAT).

One might doubt that such a complex international organization would work at all. It is truly remarkable that it works effectively and efficiently, and profitably too. As a technological organization it has made substantial progress in engineering development; as an operating entity it provides a reliable global telecommunications service with the capacity to operate effectively with national and bilateral systems; and as a commercial organization it has managed its affairs prudently and profitably.

The 11 original members of INTELSAT have grown to 95. The members and participants include nations of every political stripe and economic system. It is interesting to contemplate the diverse cultures tied together through a web of electromagnetic links. For example, one U.S. station at Etam, West Virginia, connects directly through an Atlantic INTELSAT IV-A satellite to 31 other earth stations located in such geographically and politically diverse locations as Asadābād, Iran; Cairo, Egypt; Cheia, Rumania; Lurin, Peru; Emeq Ha'ela, Israel; Lvov, U.S.S.R.; Nsele, Zaire; Pretoria, South Africa; Sintra, Portugal; and Tanum, Sweden.

The INTELSAT system keeps its books in a manner similar to any commercial organization (*9*). In principle, members of INTELSAT (countries) own investment shares in proportion to their current usage of the system. These shares range from 0.05 to 31 percent, the U.S. share. During its 12-year history (1965 through 1976), the total investment has been $650 million. With current annual operating revenues of $140 million, INTELSAT meets its financial goal of a 14 percent return on invested capital. Through technological progress and economies of scale, circuit charges have undergone a steady reduction from the unit charge (per half-circuit-year) of $32,000 in 1965 to $7,380 in 1977.

The total investment in all the earth stations of the network, owned not by the INTELSAT organization but by individual countries, is approximately that of the space segment, about $750 million. Cost analysis has indicated that the total annual half-circuit costs in the INTELSAT system, including earth station amortization and operating expenses and space segment charges, may be less than $10,000 (*10*).

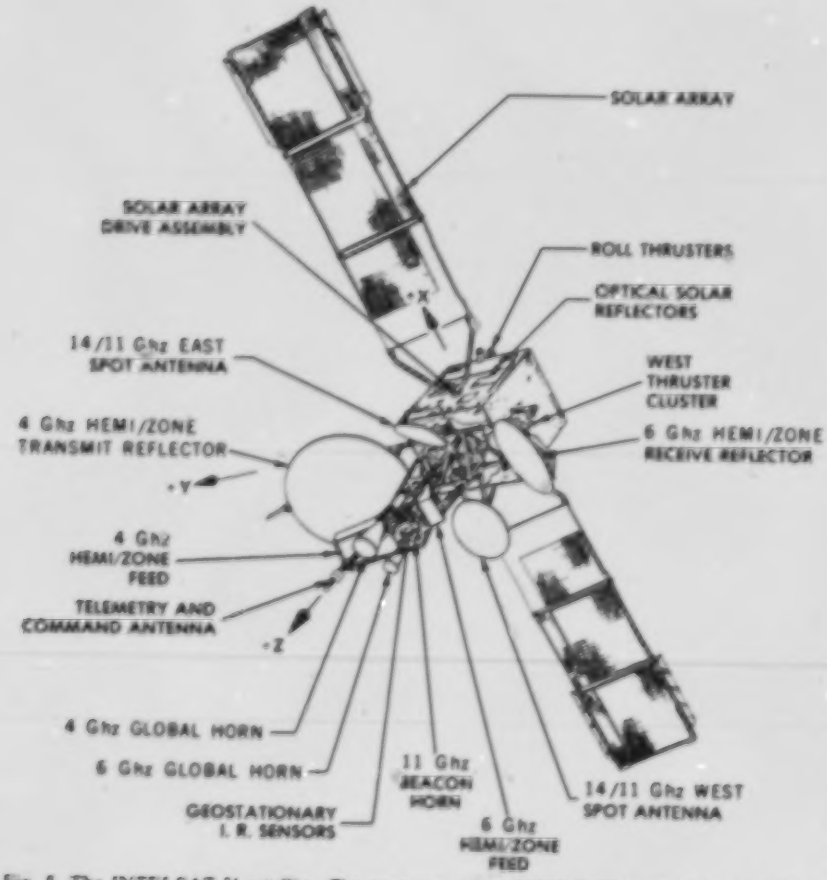


Fig. 5. The INTELSAT V satellite. The next-generation INTELSAT satellite will be body-stabilized. Like the earlier INTELSAT satellites, it will operate on the 6/4-Ghz band; it will inaugurate the use of the 14/11-Ghz band. The first launch will be in 1979 and the satellite will carry the projected traffic through the mid-1980's.

Domestic Systems

The same technologies that made it possible for satellite communications to span oceans is equally applicable to the crossing of geographical barriers or great distances within national boundaries. However, in the United States the sense of national purpose and unified action that gave impetus to the development of the INTELSAT system was lacking to the development of a domestic system. Also, there were existing competitive terrestrial communications systems. Actually, several possible routes might have been taken, either to provide domestic service through INTELSAT, or through a single existing organization or some new organization, or through competitive means. After a series of government studies and regulatory rulings lasting over a 5-year period, the United States finally adopted an "open skies" policy which led to the development of several competitive systems. While deliberations took place, some domestic service between remote areas was provided through INTELSAT between the contiguous 48 states and Alaska, Hawaii, and Puerto Rico, but these services are now being assumed by other domestic systems.

Without internal conflict, Canada established its own domestic system, TELESAT, with the launch of its first satellite, Anik, in 1972. At present, three satellites of the first generation are in orbit providing service to over 50 earth stations, many in small remote communities. Indonesia established a similar separate domestic system in August 1976.

Other countries, for example, Algeria, Brazil, Nigeria, and Norway, are leasing transponder capacity from INTELSAT for domestic service. Such services can be instituted quickly and operate efficiently since no capital investment is required for the satellites, and the country involved can concentrate on building and operating the earth stations.

The United States now has three separate domestic systems in operation. Western Union was the first of these, inaugurating service via its WESTAR satellites in 1974. Then RCA followed with SATCOM late in 1975, and, by July 1976, the COMSTAR system became operational with COMSAT General leasing

satellite capacity to American Telephone & Telegraph and General Telephone & Electronics.

Domestic systems provide communications to regions, cities, and, in some cases, individual users. These services demonstrate the inherent flexibility of satellite communications since specially tailored earth stations are used to fit specific needs. The COMSTAR system is used with 30-meter stations to serve major regions of the United States by linking telephone networks much as in the global system. The WESTAR and SATCOM systems use 10-m stations for intercity communications and television distribution but may use 5-m stations to meet special customer needs.

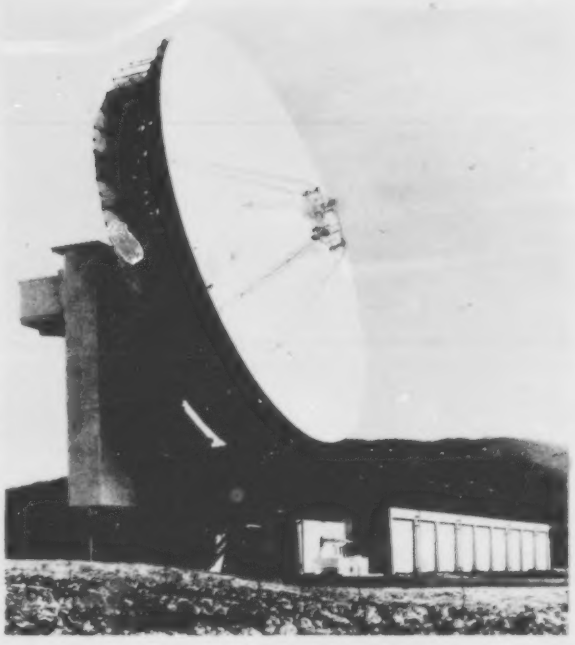


Fig. 6. Standard earth station (INTELSAT system). This parabolic dish antenna (29.5 m in diameter) is the dominant structure at the earth station for commercial satellite communications located at Jamesburg off the upper Carmel Valley in central California. The facility is capable of handling and processing all types of overseas communications: multichannel telephone, teletypewriter, facsimile, data, and both black-and-white and color television.

Mobile Systems

The use of mobile terminals is an important advantage that satellite communications has over all other modes of long-distance wide-band communications. Not surprisingly, a number of satellite systems to serve mobile platforms have been proposed, with one now fully operational. The MARISAT system, operated by the COMSAT General Corporation with several other U.S. carriers, provides three-ocean service to ships at sea. Starting in 1976, the MARISAT system (*11*) has been providing leased service to the U.S. Navy in the military very-high-frequency (VHF) band (225 to 400 Mhz) and commercial service at L band (1.5 Ghz). Thirty commercial ships flying flags of 12 nations were equipped for service with the MARISAT system at the start of 1977. This number is expected to grow to several hundred in the next few years. MARISAT provides two-way, ship-to-shore voice and telex service interconnected to the U.S. national and the international public-switched telephone networks. At present, the service cost is several dollars per minute, slightly more expensive than but still competitive with the much less serviceable and reliable high-frequency communications system which is the only other present mode of long-distance maritime communications.

What is spoken of as aeronautical, or more accurately, air traffic control communications is normally considered a governmental service. Plans for satellite systems to provide such service have been considered and proposed for a decade, but private investment capital was not forthcoming for a system dedicated to public service. However, in 1974, a tripartite organization was formed by the COMSAT General Corporation, the European Space Agency, and the Canadian Department of Transportation to develop AEROSAT, a satellite communications system to serve aircraft in international flight (*11*). The AEROSAT system would use the 136-Mhz VHF band as well as L-band frequencies close to the maritime band for this service. The U.S. customer for such service would be the Federal Aviation Administration. Difficulties in arrangements between all the parties involved have slowed progress so that, at this writing, it is difficult to predict a service date, but 1980 seems feasible.

Electronics Contributions

The technology on which progress in satellite communications has depended has come from two sources: (i) space technology, largely the product of the space program of the 1960's, heavily and continuously subsidized by the efforts of the National Aeronautics and Space Administration (NASA) and the Department of Defense, and (ii) communications technology, dependent largely on commercial development.

The developments in electronic devices and components stemming from line-of-sight microwave technology have contributed in many respects to the rapid technological progress made in satellite systems over the last decade. Several developments in electronics technology have been particularly important in contributing directly to progress in communications satellites. These include solid-state devices such as transistors, TWT amplifiers, and lightweight antennas and feed systems. Satellite electrical power systems have benefited from advances in silicon solar cells and lightweight, high-energy-density storage cells.

A typical satellite transponder is shown in Fig. 7. Early satellites, such as

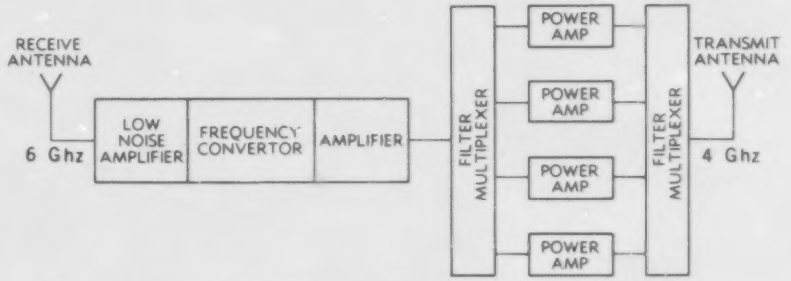


Fig. 7. Typical satellite communications transponder. A typical transponder includes a receiver incorporating a low-noise amplifier, frequency translation, and filter multiplexers; the output amplifiers raise the level of the output signal to approximately 5 to 10 watts.

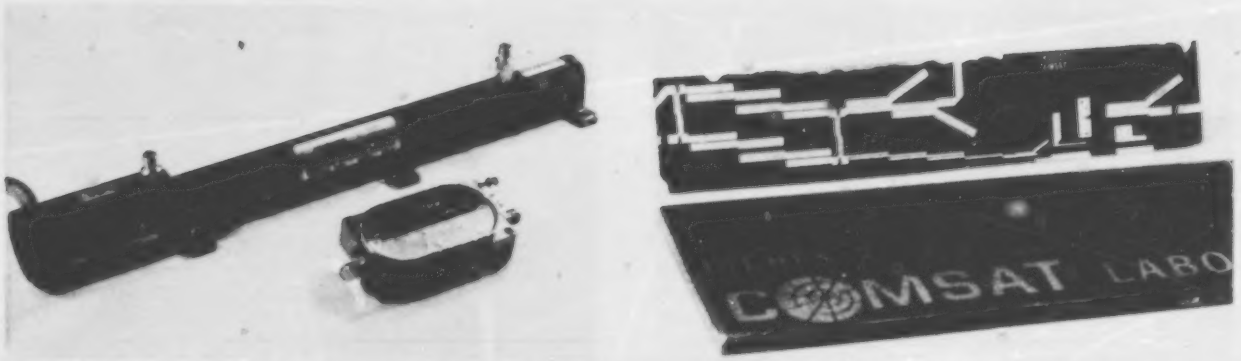


Fig. 8 (left). A modern traveling wave tube (TWT) amplifier and a solid-state amplifier. Future satellites will use solid-state amplifiers instead of TWT's, resulting in increased reliability, lighter weight, and improved transmission performance. Fig. 9 (right). Microwave integrated circuit (MIC) mixer. This MIC uses a thin quartz substrate coated with a layer of gold. The gold on one side is photoetched to a pattern which yields resonant circuits in the microwave frequencies. This modern form of construction affords the advantages of lighter weight and higher reliability, achieved as a result of a much simpler mechanical form.

INTELSAT II and III, used only two receivers and transmitters to amplify the two-way signal flow. The INTELSAT V satellite will include seven operating receivers and eight spares, along with 27 output amplifiers and 16 spares. There is a need for improved device and component reliability and conversion efficiency from primary to radio-frequency (RF) power as well as reduced weight, so that capacity can be increased. The TWT was first announced in 1947 (*12*), but it was not until the mid-1950's that it came into wide use in terrestrial microwave links. In the last decade, improvements in the TWT have resulted in an increase in the conversion efficiency from d-c to RF power from 20 to 40 percent (saturated operation) and the efficiency of the associated power supplies has improved from 60 to 87 percent. Permanent magnet-focusing based on the use of samarium cobalt magnets has further reduced the weight and improved the efficiency of these high-gain amplifiers. Figure 8 illustrates a modern TWT amplifier alongside a solid-state amplifier.

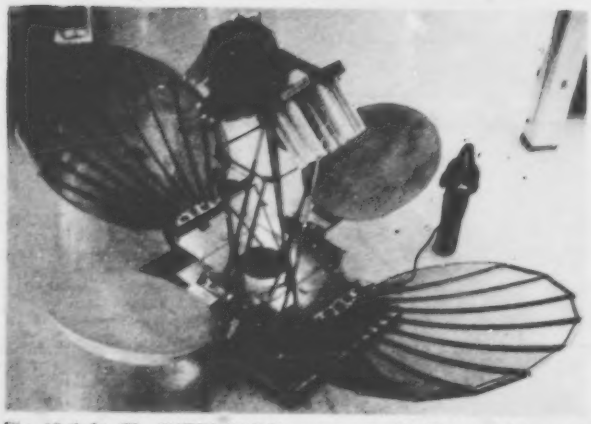
Future satellites are expected to make use of silicon power transistors to replace TWT's. These transistors operate at lower voltages (24 to 48 volts) with improved reliability, transmission performance, lighter weight, and very much longer lifetimes. At higher frequencies, IMPATT (impact avalanche and transit time) diodes can provide the power amplification that is required. Amplifiers based on the use of these devices have been built (*13*) at 11, 14, 20, and 30 Ghz, the bands assigned to space communications for future expansion.

Germanium tunnel diodes have been used as input amplifiers in the receivers of INTELSAT satellites. Low-noise transistor amplifiers, such as silicon bipolar or gallium arsenide field-effect transistors, are now replacing the tunnel diodes. These transistors can achieve a noise temperature about a third of that of the tunnel diode amplifier at the 6-Ghz receive band used in present INTELSAT satellites. For future satellites operating at higher-frequency receive bands, 14- and 20-Ghz low-loss mixers or a new invention, active medium propagation (AMP) amplifier (*14*), will probably be used. This AMP device developed at COMSAT Laboratories achieves low-noise gain through negative differential conductivity along a transmission line formed by a conductor deposited on a gallium arsenide substrate. This device has achieved a noise figure of 10.7 db at 16 Ghz.

To avoid interference to the input signal of the satellite caused by the retransmission of the signal back to earth, the receive and transmit frequencies must be separated. The amount of frequency separation is dependent upon the isolation that can be achieved by input filters and transmitting and receiving antennas. Typically an isolation that is 40 db more than the maximum direct gain between transponder input and output is required (that is, about 120 to 150 db). The frequency translation is obtained when the incoming signal is mixed with a locally generated signal. Coaxial lines, used as the mixers in INTELSAT IV, achieve a conversion loss of 7.5 db and weigh approximately 360 g. Designs in which microwave integrated circuits are used have realized less than a 4-db conversion loss and weigh about 75 g (*15*). The microwave integrated circuit (MIC) (Fig. 9) uses a thin quartz substrate coated with a layer of gold several micrometers thick.

Complex circuits are physically realizable through computer-aided design, and more stringent specifications, in terms of loss and gain variation within an operating band, are being achieved. The local oscillator part of the frequency translator must have a long-term frequency stability usually better than one part in a million per year and little jitter, to avoid adding noise to the amplified signal. In the standard approach a quartz crystal is used to control the oscillator frequency at 15 to 100 Mhz. Amplifiers drive a transistor multiplier to approximately 2 to 3 Ghz. To obtain higher frequencies, varactor frequency multipliers are used. Here again, improvement in the efficiency of the power conversion from d-c to RF and reduced noise are being realized with a transistor oscillator operating directly in the gigahertz range and stabilized by a highly selective microwave cavity.

Very dramatic advances have taken place in the art of antenna design. Early satellites, mentioned earlier, used simple radiators which spread the microwave energy over a wide solid angle very much larger than that subtended by the earth's disk. With the improvement in spacecraft materials along with more precise control of the spacecraft attitude, the beam width of antennas has been decreased and the energy concentrated toward the area of the earth being served. In addition to increasing the effective radiated power, the narrow beams permit the frequency bands assigned to the satellite to be reused. Figure 10 illustrates an antenna system which will generate the beam patterns specified for the INTELSAT V satellite. Separate beams directed at the Eastern and Western hemispheres use one polarization and zones coincident in the Eastern and Western hemispheres use the orthogonal polarization at 6 and 4 Ghz, thus realizing a fourfold reuse of these frequency bands. In addition, two beams at 14/11 Ghz will provide coverage in the



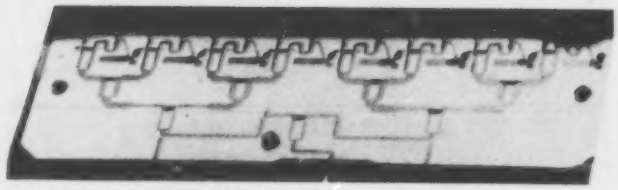


Fig. 10 (left). The INTELSAT V antenna model. This multibeam antenna model, constructed by Lockheed Missiles & Space Corporation under an INTELSAT research and development contract, will generate the beam patterns similar to that specified for INTELSAT V. The antenna system consists of two 3-m unfurlable offset-fed parabolic reflectors for the 4-Ghz coverage, one 6-Ghz offset-fed parabolic reflector, and one 14/11-Ghz offset-fed reflector. Fig. 11 (right). One section of a 16 by 16 MIC switch and the controlling LSI unit. This type switch will provide interconnection of the many narrow antenna beams that are expected to be used on future satellites.

areas with high traffic density in northwestern Europe and in northeastern United States and Canada. The total bandwidth available will be about 2300 Mhz (including guard bands between transponders). The use of lightweight material, exhibiting high strength and a low temperature coefficient, in which epoxy-filled carbon fibers form the major structural elements of the antenna and metallized Dacron mesh forms the reflecting surfaces has made such complex antenna systems feasible. Multielement feeds, such as those shown in Fig. 10, have been designed to achieve complex coverage.

Future satellites are expected to use a great number of very narrow beams, of the order of a few tenths of a degree, essentially one directed to each earth station. These beams must be interconnected. The interconnection is achieved through microwave switches operating at approximately the 4-Ghz band or, alternatively, at base-band frequencies. The microwave interconnection is effected by switches that use semiconductor diodes. These are controlled by an on-board microprocessor which cycles the connection of the receivers and transmitters to the several antenna ports so that all stations are cyclically interconnected to each other. Figure 11 shows one section of a 16 by 16 switch and the LSI (large-scale integration) unit controlling the MIC crossbar switch.

Further developments in microwave integrated circuits and transistor devices will allow a dramatic reduction in the size and weight of receivers which can demodulate digital transmissions, and, with the availability of large-scale integrated circuits for base-band processing, interconnection at the base-band signals is becoming feasible. This base-band processing will require that the system modulation be digital and the access to the satellite be by time interleaving. The interconnection at base-band also permits the processing of signals so that less power and bandwidth is consumed per bit of information. Thus, the combination of large-scale integrated circuits at lower frequencies with microwave integrated circuits at gigaher z frequencies will make possible very much more complicated satellites with capacities approaching 50,000 to 60,000 telephone circuits.

System Trends

The rapid technological progress in electronics coupled with the impressive operational results and economic returns over the past decade have spawned new services, new systems, and new concepts for satellite communications. A number of experimental satellite programs will provide a good basis for further technological advancement, which will, in turn, provide new services and systems. The last U.S. experimental communications satellite was ATS-6, launched in 1974, which among other accomplishments demonstrated television broadcast to the Indian subcontinent. Other countries followed: Canada with the Communications Technology Satellite and France and Germany with Symphonie. Italy expects to launch the Sirio experimental satellite and Japan two experimental communications satellites in the next year; NASA has not pressed development work in communications satellites for several years but has recently considered resuming a major role and has been urged by several professional societies and the National Academy of Engineering to do so. It is likely that NASA will support extensive satellite communications experimentation in the flight test program for the Space Shuttle.

The Shuttle will open new avenues for the design of communications satellites. Shuttle-launched spacecraft can be much bigger and heavier than current designs with minimum cost penalty. Larger satellites in geostationary orbits can be used as stable platforms on which to erect a large and complex antenna array to promote the use of multiple frequencies and multiple beams—possibly multiple-service satellites as well.

Communications satellites of the future are likely to make use of the on-board processing mentioned above. The satellite will then carry out signal-processing functions including signal reshaping, switching or remultiplexing, or both, and compression. Operating at digital data rates above 100 megabits per second, these satellites will require very reliable digital logic and microwave integrated circuitry.

In addition to the 6/4- and 14/11-Ghz bands, where 500 Mhz of bandwidth is available in each band, the next decade will see the multiple reuse of these bands. For very-high-capacity links, the 30/20-Ghz band, where 3500 Mhz of bandwidth is available, will be coming into use.

Earth stations in future systems will be smaller, less costly, and closely adapted to the users' needs. Earth stations intended to serve cities, towns, villages, or individual users will be sized, configured, and located differently from those designed to serve countries. Earth

stations to serve users in ships, aircraft, and land vehicles (land-mobile) will be designed differently and specifically for the use they serve (see Fig. 12).

Earth stations will be manufactured in large quantities to be economical. Those with antennas of 5 m or less are already priced below $100,000, and the cost of small receive-only stations is expected to approach $10,000 in the next decade.

The trend to digital transmission in satellite systems will in some cases precede the trend to digital transmission in terrestrial systems for the reason that the new satellite systems can employ new technology faster than older systems that involve replacement of existing plant facilities. Advances in solid-state technology, particularly the availability of reliable high-speed logic, have made the time domain easier to work in and more flexible than the frequency domain. The advantages of digital transmission techniques include efficiency and economy in switching, signaling, processing for compression and error control, modulation, and multiple access. Digital transmission methods will be used to carry traffic originating in digital form (for example, computer data and facsimile) but also to transmit more efficiently traffic originating in analog form (for example, voice and television). Domestic and special-purpose systems can be expected to use primarily digital transmission techniques. The TELESAT system of Canada introduced the first commercial operation of time division multiple access (TDMA), a digital technique that in-

Fig. 12. Small earth stations. An inherent advantage of satellite communications is the wide area coverage, including the ability to reach as close as desired to the ultimate user of the service provided. Smaller, inexpensive special-purpose earth stations are frequently required for such service. (A) An earth station equipped with a 10-m antenna used to provide domestic telephone and television service to a remote community in Alaska. Similar stations are now being used at some INTELSAT locations. (Photo courtesy of Scientific-Atlanta, Atlanta) (B) An antenna (5-m diameter) next to a Dow Jones printing plant used for high-resolution facsimile transmission of *The Wall Street Journal*. (C) An oil tanker with a mast-mounted 1.2-m antenna (inside the white radome) which provides a link to the shore communications networks via the MARISAT satellite. (Photo courtesy of COMSAT-General Corporation) (D) A 1.2-m antenna which can transmit and receive in the 14/12-Ghz band. This station was developed by COMSAT Laboratories for the demonstration of emergency communications with the American Red Cross.

volves transmission in short bursts and allows several earth stations to communicate sequentially on the same frequency. The second system to use TDMA was MARISAT, for ship-to-shore telex. Future systems, such as that proposed by Satellite Business Systems, a new company formed by COMSAT, International Business Machines, and Aetna to enter the U.S. domestic field, can be expected to be all digital.

The INTELSAT system, although probably best known for its introduction of international television, still derives over 90 percent of its revenue from telephone service. The future holds promise for as great an international market for data, including graphics—facsimile or "electronic mail"—and the evolving new field of "computer networking."

The U.S. domestic market for data transmitted by satellite is much larger than that for voice, since the public-switched telephone network in this country is already so extensive and efficient. For high-speed data the switched network is not now adequate, nor with its multiple cascaded links will it ever be as attractive as satellites from an error performance standpoint. In the United States, certainly, the television market represents an excellent prospect for satellite transmission. Also, new services such as video-voice teleconferencing may become practical on a large scale through domestic satellite systems. One can foresee the application of all these same service possibilities—voice, video, and data—to different degrees to mobile systems (maritime, aeronautical, and land-mobile).

Finally, satellite communications hold great promise for collective services, that is, service to great numbers of users, usually small, of a nature other than point to point. Satellites, because of the wide area covered, are ideal for providing multiple-origin or multiple-destination service. The former, generally known as "data collection systems," are applicable to the transmission of small quantities of data from numerous points of origin to a central delivery point for processing. The military services use such systems for warning, intelligence, and surveillance. Civil needs for meteorological, climatic, agricultural, geographic, and geological data are abundant. Most of these services are traditionally provided by government, but there are many possibilities for commercial applications.

Point-to-multipoint or distribution service may also be provided ideally by satellite. One such service, already established through domestic satellite systems, is television distribution—the satellite and the cable being a natural combination for the distribution of television. Over 40 cable television locations in the United States are now served through the RCA SATCOM system.

Another distribution service already being provided is high-resolution facsimile. *The Wall Street Journal* appearing on desks in New York and Miami is transmitted by satellite from the Dow Jones composing office in Chicopee, Massachusetts, to their printing plants in Brunswick, New Jersey, and Orlando, Florida, respectively. It is anticipated that additional "satellite" plants will be added across the country to make the *Journal* a truly national newspaper.

Distribution satellites providing both television and daily newspaper service across the country—and across the oceans—can certainly be said to have reached the "peoples of the world." Whether these services will help to bring "peace and closer brotherhood," as President Kennedy put it, has yet to be determined.

References and Notes

1. J. F. Kennedy, "Policy Statement on Communications Satellites," 24 July 1961, Report to the Congress from the President of the United States, U.S. Aeronautical and Space Activities (Government Printing Office, Washington, D.C., 1961), pp. 105–106.
2. A. C. Clarke, *Wireless World* **51**, 303 (October 1945).
3. J. R. Pierce, *Jet Propul.* **25**, 153 (April 1955).
4. L. Jaffe, *Astronaut. Aerosp. Eng.* **1** (No. 8), 48 (September 1963); J. R. Burke, *ibid.*, p. 72; S. Metzger and R. H. Pickard, *ibid.*, p. 64; I. Welber, *ibid.*, p. 68; J. R. Pierce, *The Beginnings of Satellite Communications* (San Francisco Press, San Francisco, 1968).
5. U.S. Congress, *Communications Satellite Act of 1962*, P.L. 87-624, 87th Congress, 31 August 1962.
6. B. I. Edelson, R. Strauss, P. L. Bargellini, *Acta Astronaut.* **2**, 691 (1975).
7. E. T. Jilg, *COMSAT Tech. Rev.* **2** (No. 2), 271 (fall 1972).
8. R. J. Rusch and D. G. Dwyre, paper presented at the 27th International Astronautical Congress, Anaheim, California, 10–17 October 1976.
9. International Telecommunications Satellite Organization, "Annual Report to the Secretary General of the United Nations by the Executive Organ of INTELSAT," 1975–1976 (INTELSAT, Washington, D.C., 1976).
10. B. I. Edelson, H. W. Wood, C. J. Reber, *Raumfahrtforschung* **20** (No. 2), 83 (March–April 1976).
11. W. Luksch and E. J. Martin, paper presented at the 27th International Astronautical Congress, Anaheim, California, 10–17 October 1976.
12. J. R. Pierce, *Proc. IRE* **35**, 111 (February 1947).
13. M. Barrett, *COMSAT Tech. Rev.*, in press.
14. P. L. Fleming, *Proc. IEEE* **63**, 1253 (August 1975).
15. L. Pollack, in *Proceedings of the 1971 European Microwave Conference* (Royal Swedish Academy of Engineering Sciences, Stockholm, 1971), vol. 2, p. C 11/S.1.

Computers and Research

W. O. Baker, W. S. Brown, M. V. Mathews, S. P. Morgan,
H. O. Pollak, R. C. Prim, S. Sternberg

Overview

The revolution of the research process itself by "computers"—foreseeable for 20 years, discernible for 10 (*1*)—is now well advanced. Computers have been adopted, adapted, and absorbed into every aspect of the research process, with aggregate consequences that are pervasive and profound.

Twenty years ago, Bell Labs Research had acquired its first modest (IBM 650) computer. A mere handful of mathematical enthusiasts, privy to the arcana of transforming equations into effective sequences of machine code, were stimulating applications throughout the research area. But "computing" was, figuratively and literally, only a minor annex of mathematics research. Today, large modern computer centers provide efficient, scientist-oriented batch and time-shared services. And throughout the research areas are scores of dedicated minicomputers and hundreds of interactive terminals—an average of one mini and five terminals per 15 professionals. First-hand experience in computer use is the norm, and at least minimal literacy in "computerese" is nearly universal. Computer pervasion of the research process is virtually complete.

Idea generation, the crux of the creative process (Fig. 1), is influenced in fundamental ways, such as the kinds of ideas, what they are about, the terms in which they are conceived, and the stimuli to which they are a response.

Broadly, these are now more complex ideas about more complex matters. But a complicated idea is worthless unless something can be done about it. What has changed is the usable level of idea complexity. Computers have significantly expanded the domain of tractable complexity. The exploration processes through which ideas must be validated and refined into science have been augmented and expedited.

Computer models involving many elements, interrelated in specific but analytically opaque ways, can now be evaluated, manipulated, and explored in pursuit of confirmations, insights, and narrowings of uncertainties. Thus the development of computers has made possible extensive explorations, validations, and refinements that in some cases would have been literally impossible, and in others would have been addressable only through very extensive, expensive and time-consuming programs of experimentation, data collection and analysis.

Thus, in many instances, computer studies afford quick, cheap alternatives to older exploratory methods. But they are also used to facilitate, expedite, and enhance field and laboratory experimentation and demonstration by prior targeting of critical parameter combinations for exploration and of decisive variables for measurement. The productivity and effectiveness of laboratory experimentation has been further enhanced by the exploitation of computers in order to achieve "nicer" and more elaborate control, as well as quick, reliable data collection, analysis, summary, and display.

Not all potentially useful data can be acquired under tidy, controlled laboratory conditions—notably in fields such as ecology, psychology, and economics. New computational-graphical techniques now enable the profitable probing and exploration of previously intractable masses of imperfect, inadequately structured data to find useful new patterns, segmentations, correlations, trends and visualizations. Accordingly new clues to understanding, inferences of structure, and model conceptions are derived.

The permeation of the exploration phases of the research process by computer techniques feeds back upon the antecedent phases of cerebration and inspiration in significant ways. The new motivations and stimulations encourage idea conception more often in terms of specific logical, quantifiable relationships and processes, thus enhancing sharpness and coherence of thought in all sciences.

A further phenomenon is the symbiosis between the overall research process and computing science itself. Stimulated by their personal activities in the exploration phases, scientists of all kinds conceive ideas about computer use that are applicable beyond their own fields, and thus contribute directly to computing science. Furthermore, the questions and complaints of the diverse community of sophisticated scientist users provide invaluable motivation, direction, and focus for the computing research activity per se in evolving new techniques and methodologies of ever greater real utility to the research community.

In this article, we elaborate on the foregoing general observations by examples drawn from four different research sectors. First, however, we discuss important additional contributions of computers to the flow of information nurturing the research process.

Getting Information In, About, and Out

The effectiveness of the research process depends critically on scientific and technical communication, a process that is deeply and increasingly intertwined with the continuing electronic revolution. Researchers must communicate among themselves to maintain an awareness of related fields, and with others to ensure that their work will be relevant and that their results will be applied.

Perhaps the most important instrument for scientific and technical communication is the telephone, which is supported by the international telecommunications network. A trend is now evident toward providing scientists with computer terminals so that they have become able to exchange written messages with speed and convenience, and to share in the development and use of an evolving store of programs and data. Nevertheless, a research effort cannot have maximal impact until its results are formally documented, published, abstracted, indexed, and used.

The authors are members of technical staff at Bell Laboratories, Murray Hill, New Jersey 07974. This article was photocomposed, including the final page layouts, in the Bell Laboratories Computing Science Research Center, using a general purpose text formatting program called TROFF. TROFF runs under the UNIX operating system on DEC PDP-11 computers and drives a phototypesetter.

Once written information exists, scientists depend heavily on libraries to help them find it, and libraries are using computers in a growing number of ways to become more efficient and effective. For example, the Bell Laboratories Library Network (*2*) consists of 25 libraries serving more than 10,000 technical and managerial employees in nine states. To help these libraries work together as a single entity, the BELLREL system maintains an on-line file of the entire book and journal collection. The complete union catalog is published annually in book form with monthly cumulative supplements, and copies are available to provide ready access for all employees.

At Bell Labs the library's MERCURY system uses a computer to route, on the basis of subject matter and other criteria, internal technical reports to interested employees. The library also publishes a dozen current awareness bulletins, prepared in whole or in part by computer from external magnetic tapes or locally entered data, or both. For information retrieval, the library provides literature searching of commercial and governmental data bases, and often publishes the results as specialized bibliographies.

Until recently, nearly every journal article was typed several times at the author's institution, and then retyped or typeset by the publisher. Later its title, abstract, and bibliography would be retyped again and again for abstracting, indexing, reviewing, and searching. Typically each retyping would correct some errors and introduce others. Fortunately, an article can now be typed once into a computer, and edited as often as the author, the referees, and the editors require. When the article is in final form, a computer-driven phototypesetter can produce camera-ready copy that is essentially indistinguishable from the best manual typesetting. Special programs are available for mathematical equations and complex tables. Citations can be obtained from an on-line data base, and numbered automatically. Computers can also help in locating spelling errors, making notational changes, and constructing indexes. Even page layout can be accomplished by computer, thus eliminating the need for galley proofs.

As is usually the case with automation, the most important benefit of computer-aided document preparation is that it not only does previously expensive work more cheaply, but it also creates wholly new possibilities. For authors, the results of scientific computations can be fed directly into a paper, while programs, equations, and tables from the paper can be fed back to compilers, formula manipulators, and other programs for verification and further research. For publishers, the economies of total automation may at last transform the MERCURY-like personalized journal envisioned by Brown, Pierce, and Traub (*3*) into a profitable reality.

We are witnessing the birth of a new system (Fig. 2) for scientific and technical communication. As an author conducts his research, often largely inside a computer, his report will evolve with it. When the report is finished, the myriad of activities that follow—from printing and distribution to storing, searching, and retrieving—will be based on computer processing of successive versions and subsets of the original document.

Speech

Speech research provides an excellent example of the impact of computing on a research area. In the mid-1950's, computers rescued speech research from interminable instrumentation work that slowed progress to a snail's pace. As the power of computation increased, computer-aided research dominated the field so that by the late 1960's almost everything was done by program writing. In the 1970's digital processing became so powerful and cheap that computerlike devices were incorporated into speech communication systems, replacing analog devices. The research tool itself became the product of the research.

Speech processing in 1977 makes an interesting contrast to speech processing in 1957 when, at Bell Labs, the first digital speech material was stored in 30,000 computer cards, which represented all of 2 minutes of speech. The cards were processed with an IBM 650 computer. A single study of a simple encoding often required 24 hours. Extremal coding was invented in this way (*4*). The computer had only to locate successive maxima and minima of the speech waveform. Although the computer was slow, any other way of doing the task would have required a year of equipment building.

In 1977, by contrast, one of the fastest and most sophisticated speech synthesizers speaks in real time by dint of its ability to multiply and add in 250 nanoseconds and take square roots and divide in 2 microseconds. The machine is called a linear predictive synthesizer and is constructed from 700 integrated circuits. It is digital, but is no longer a stored program machine since it was built to perform one specific function. However, it was carefully simulated on a stored program machine before being constructed.

Speech research introduced a kind of simulation that provided a new way for people to interact with computers. Computers previously had been used to analyze data or to evaluate long complex expressions. To evaluate speech transmission systems, a human listener must hear the transmitted sound. No formula can adequately predict the

Fig. 1. The research process—in 1027 words.

quality of the transmission. In order to simulate a system on a computer, input speech must be converted into numbers, which then are processed by a computer. The computer calculates numbers representing output speech and these are ultimately converted to a sound wave. Thus, analog-to-digital and digital-to-analog converters were developed along with speech systems. During the development, the size of the converters shrank from hundreds of pounds of vacuum tubes costing thousands of dollars to single integrated circuits costing tens of dollars.

Thousands of numbers per second are necessary to represent high frequency signals—56,000 bits per second is the current standard for telephone speech. Processing and storing these numbers pushed the development of computers in important directions. Magnetic core memory plus digital magnetic tape data storage were the necessary ingredients. The IBM 704 and its transistorized successor, the 7094, were the workhorses of our speech simulation to the mid-1960's. After that time, big computer development turned toward satisfying multiple users with time-sharing programs. The shared machines were not appropriate for the heavy computations and real time demands of speech. Consequently, speech researchers turned to minicomputers as their principal tool. This encouraged the development (i) of cheap, small, powerful central processors, (ii) of the direct connection of computers with laboratory equipment, and (iii) of visual displays on scopes. The development of many of the features of today's minicomputers was strongly influenced by these computing demands (*5*). Eventually, networks of minicomputers emerged and pointed the way to distributed but interconnected computing power which may be the principal pattern for the next generation of computers.

Figure 3 shows a computer at which a researcher is synthesizing speech. As she hears the speech, she watches the movements of the articulators on a scope. She can record speech directly into the computer memory by speaking into a microphone. She can control the operation of the computer with buttons and knobs. The computer has much information on its disk files including past speech samples and a complete pronouncing dictionary. Long computations can be sent to a special high-speed central processor.

With such computing equipment, new ideas can be tried quickly and be improved or rejected according to their merit. Efficient speech encodings, excellent quality speech synthesis, limited automatic speech recognition, verification of a talker's identity, and progress toward a deeper understanding of the nature of speech have been achieved (*6*). The results of this research can now be directly applied, embodied in today's integrated circuits and microcomputers.

Fig. 2. Evolving system for scientific communication.

Atmospherics

Atmospheric chemistry exemplifies the power of new computer-dependent exploratory methodologies in modeling many-component, dynamic, interactive processes, and in statistically discerning useful forms and patterns in great masses of data.

Early in the revolution, modeling studies were decisive in unraveling the mysteries of the exotic phenomena observed in hypersonic reentry of space vehicles. Here the computer, being given chemical reactions, generated internally differential equations and solution schemata. Another example is the priority study of fluorocarbon reaction dynamics in the upper atmosphere.

Fig. 3. Computer speech synthesis from text.

As to the lower atmosphere, a timely and exciting example is afforded by the first comprehensive study of the urban air (*7*). This study has demanded and strained the capability of the computer to handle simultaneously a complex chemical model of reactions in mathematical form, a large data base of atmospheric measurements, an extensive library of robust methods of data analysis, and a variety of graphical displays. The results have demonstrated the power of the use of chemistry, mathematics, and statistics together with the computer to provide incisive contributions to the solution of controversial and difficult problems.

The aim is to understand the scientific basis for contamination of the urban atmosphere. Thus the first requirement is to represent to sufficient accuracy the chemical kinetics of the lower atmosphere. The associated differential equations are stiff, but recent mathematical developments have made them amenable to computation (*8*). The model contains 70 distinct chemical species and 130 reactions among these species, each with its own characteristic constants. Apart from the variety of individual reaction rates, other essential factors include the influence of sunlight, both through photochemical reactions and through the height of the mixed layer of the troposphere, air turbulence, and bulk air flow.

The essential novelty of the approach is not primarily in the size of the mathematical system—roughly twice the number of variables and equations in previous models—but especially in the successful modeling of sulfur and of aerosol reactions, and in the fact that the constants associated with each reaction were taken from previous theory or experiment rather than iterated for an overall best fit. The latter decision produced greatly increased confidence in the model at the expense of previously unmanageable computational complexity.

The data base contains more than three million observations from approximately 40 primary and 60 secondary locations in the band from Pennsylvania to Massachusetts. They include a great assortment of air quality

measurements and meteorological observations. Differences in instrumentation and measurement techniques at the various stations, missing values, and occasional poor or extremely deviant observations affect 5 to 10 percent of the data, and make it vital that the statistical analyses be insensitive to such disturbances. These requirements made the project a critical but grateful customer for the developing arsenal of robust methods of data analysis.

Computer generated displays of results, at their best, give insights with an economy and power not easily obtained otherwise. They facilitated the accomplishment of the dual aims in this study of the urban atmosphere: to test the agreement of the complex chemical model with the observed facts, and to display the insights produced by the robust statistical analyses. Three recently developed tools of especial value are quantile probability plots, plots of moving statistics, and box plots. The quantile-quantile (Q-Q) plot displays quantiles of one distribution against those of another, and permits graphical comparison of two cumulative distribution functions without any prior assumptions regarding the nature of these distributions. It is especially useful in comparing workday with Sunday observations of various contaminants. Moving statistics superimposed onto a two-dimensional scatter plot provide local summaries of data whose structure is difficult to perceive otherwise. Such plots, for example, permit us to see characteristics of the distribution of contaminant concentrations as a function of wind direction. Thus the importance of the transport of pollutants by wind currents becomes evident. A box plot allows simultaneous presentation of extreme high and low values, upper and lower quantiles, and the median for an ordered series of distributions—such as monthly concentrations of various pollutants.

The study has produced an atmospheric chemical model of the New Jersey–New York area which shows good agreement between predicted atmospheric contaminant levels and those determined experimentally from field data. Two sample results from the study (*7*) are as follows.

1) Aldehydes, which themselves promote formation of other contaminants, are produced chemically in the atmosphere besides being emitted from motor vehicles.

2) While most primary emissions, such as carbon monoxide and nitric oxide, are noticeably lower during Sunday morning hours than on workdays, ozone levels are at least as high on Sundays as on workdays (see Fig. 4) (*9*).

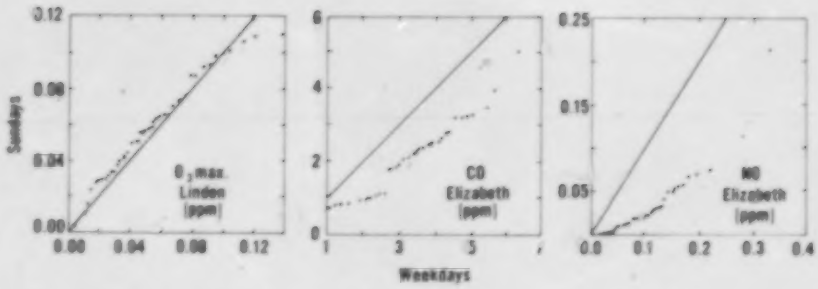


Fig. 4. Quantile-quantile plots for Sundays as compared to workdays.

Ozone is frequently considered a representative atmospheric contaminant, and tends to dominate air quality standards. The combined power of chemistry, statistics, and the computer has enabled us to understand that the actual ozone levels on any given day are the result of a delicate balance between several major atmospheric processes.

Psychology

In psychological research the computer revolution has influenced both method and theory. In perception research, for example, the computer generates and permits precise control over complex visual and auditory patterns; in psychophysiology it provides a rapid high-capacity storage medium for multichannel analog responses such as electrical brain waves; and in sensory and learning studies it increases experimental efficiency by introducing flexible real-time management of adaptive testing, in which prior responses guide the choice of new observations. Laboratory computers have accelerated research progress and encouraged exploration by facilitating changes in experimental conditions and reducing—sometimes dramatically—the delay between data collection and analysis. Without computers, some of the widely used methods of multivariate statistical analysis would be impractical because they require iterative operations on large numbers of observations. One such method is multidimensional scaling, which reveals patterns in complex psychological data by permitting diverse sets of stimuli to be represented as points in a space whose dimensions correspond to the attributes that determine how the stimuli are perceived or judged (*10*).

Computer science has also expanded and altered the kinds of questions psychologists ask and the nature of the acceptable answers (*11*). During the first half of this century, scientific psychology, dominated by behaviorism, focused mainly on the effects of repeated trials on the directly observable relations between stimuli and responses, and hardly dealt with describing the perception, memory, and thought processes that intervene. This stemmed at least partly from the lack of an adequate formal language for characterizing such processes; whereas the available mathematics was too restrictive, ordinary prose was too vague. Computer science, however, made new languages available and persuaded psychologists that programming concepts might be acceptable as precise descriptions of information processing in man as well as machine. Moreover, the hardware-software distinction supported the legitimacy of theories couched in terms of abstract information-processing operations in the mind rather than in terms of neurophysiological or chemical processes in the brain. A new field of cognitive psychology flourished as psychologists realized that concepts such as "feature extraction," "coding," "subroutine," "search," and "parallel processing" are ideally suited for theorizing about higher mental processes. In provocative new models, complex activities of thought and decision, whether conscious or unconscious, are conceptualized in terms of input, transformation, storage, retrieval, and comparison of internal representations—operations that are concatenated in systematic flow charts. Some investigators have gone even further, stating their models in actual programming languages and then testing them by comparing computer simulations with real data (*12*).

The new interest in the constituents

rather than merely the consequences of mental activity has led to a revival of reaction-time methods. In turn, these methods provide some of the strongest supporting data for a flow-chart concept of the mind by showing that the time between stimulus and response can be a *sum*, composed of the durations of a series of separate *stages* of processing (*13*). An example is the discovery that memory retrieval is sometimes accomplished by internal serial search processes analogous to ones that a computer might use. In Fig. 5, the bottom function shows the time to decide whether a test digit was present in a previously memorized list of randomly ordered digits. Although the decision seems automatic and instantaneous, the time it takes to make the decision increases linearly with the length of the list, indicating a serial search. The top function indicates a slower search to ascertain where in the list a test digit was located.

The increasing use of reaction-time measurements is one of many consequences of psychologists' new view of man as an active processor of information, characterized by speed and capacity limits.

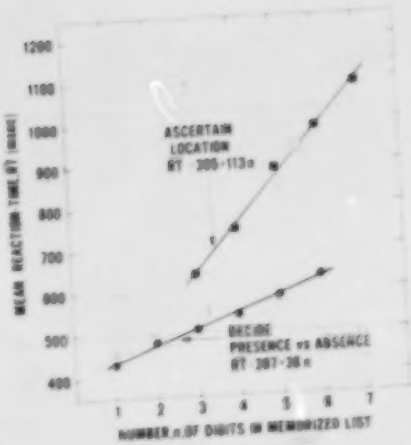


Fig. 5. Examples of experimental results suggesting that mental processes might be usefully regarded as concatenations of elementary information-processing operations.

Computing Science

Computing science both supports and stimulates the research process, by providing tools and techniques that make it ever easier for researchers to do what they find useful, and by probing and extending the boundaries of what is computationally possible.

Beginning as a subfield of applied mathematics, computing science now exerts a profound influence on almost all parts of its parent discipline. Computers have led to new kinds of attack on many classical numerical and statistical problems. Computing science itself has raised fundamental questions about computational complexity (how does the time required to solve a problem of size n by a given algorithm grow with n?)—questions that continue to occupy many first-class mathematicians (*14*). Complexity theory, whose contributions thus far range from the fast Fourier transform to proofs that a wide range of combinatorial problems are in some sense equally intractable ("NP-complete"), has changed the way people think about algorithms, and has led to new emphases on fast heuristic procedures for "finessing the impossible," that is, getting reasonable approximate solutions where exact solutions are known or conjectured to be unacceptably lengthy.

Computing science has also done much for people who merely want to use computers to "do their thing." Fortran is still the workhorse of scientific computation, but it has many disadvantages. The available control instructions in Fortran lead to complicated logical flows, often difficult to unravel a few months later. Furthermore programmers tend to use the special features of their local compiler, which plays hob with portability. Finally, Fortran is not adapted to systems programming if one wants to manipulate individual bits and bytes.

Many new tools have resulted from the theory of programming languages. Fortran preprocessors (*15*) have been designed, themselves written in Fortran, which provide additional control instructions leading to a much simpler, "top-down" logical organization for every program. Verifiers (*16*) exist that will take any program and automatically flag every instruction that does not conform to a standard subset of Fortran, thus essentially solving the portability problem.

The creation of new languages tailored to specific tasks has been facilitated by the development of practical compiler-compilers—programs that accept an abstract description of a language and generate a compiler for the language on a given computer. With a compiler-compiler it is possible to design a special-purpose language, refine it, and get an acceptable human-engineered final product within a matter of weeks.

Modern time-shared operating systems, such as Bell Labs UNIX (*17*), provide a well-designed file system for storing memorandums, programs, and data. The user enters alphabetical or numerical characters, typically via a keyboard terminal, and gives the file a name, and the system automatically stores it. To get the file back, one merely types in its name and the output device or program where it is wanted. The intended use does not have to be known when the file is created.

Under UNIX, numerical or textual data may flow through a whole sequence of programs without the user having to store intermediate results. This encourages people to produce small programs that perform a single intelligible function, and then hook them together in different ways. The building of such tools enables programmers "to stand on each other's shoulders rather than each other's feet."

A flexible file system and convenient facilities for entering, revising, concatenating, and running programs make a system like UNIX useful as a "workbench" for preparing and debugging programs, which are then run either on UNIX or on other computers via data links. Simultaneously UNIX keeps track of program documentation and of message traffic among its users. In such a research environment each person's contributions, if they meet a general need, rapidly get used by others. The result is a substantial cumulative and synergistic effect.

Scientific computing used to be mostly batch-oriented number-crunching. With the advent of simple, elegant operating systems and computer-communications networks, a user can now reach a whole federation of large and small computers, general-purpose and specialized, through a keyboard or display terminal. One can communicate and collaborate with other users, program in a variety of user-oriented languages, have jobs run on whatever machine one selects, and present the results in visible, audible, graphical, or printed form as desired. The interaction among computing scientists, mathematicians, and the rest of the research community has made an ever larger number of scientists effective users (and knowledgeable critics) of the increasingly powerful tools provided by computing, to the benefit of the whole research process.

Conclusion

Our discussion has of necessity been limited to the research process alone. However, we might have pursued the evolution of new science beyond research into the engineering and development processes, and cited counterparts of the same revolutionary trends and effects. Even within research, we haven't even alluded to whole broad areas of physics, biology, materials science, mathematics, economics and systems science from which could have been drawn other striking instances of revolutionary computer impact through enhanced mastery of complexity. Our case has been understated.

What of the future? The revolution, well begun, has by no means run its course. While no one can foretell its form, some of the propelling forces are clear. The hardware costs of computing power continue to drop by two orders of magnitude per decade. Great strides are being made in facilitating and automating the production of software for complex procedures. Methodology is evolving for the aggregation of numerous, specialized processing units into distributed, federated entities with new orders of capabilities. And, at a most fundamental level, new bridges of common technique, understanding and philosophy are abuilding between and among the diverse disciplines—deriving from a shared common language for discourse on complexity. Computerese is the new *lingua franca* for science.

References and Notes

1. W. O. Baker, *Daedalus* **99**, 1088 (1970).
2. R. A. Kennedy, in *Industrial Information Systems*, E. B. Jackson and R. L. Jackson, Eds. (Dowden, Hutchinson and Ross, Stroudsburg, Pa., in press).
3. W. S. Brosch, J. R. Pierce, J. F. Traub, *Science* **158**, 1153 (1967).
4. M. V. Mathews, *IRE Trans. Inform. Theory* **IT-5**, 129 (1959).
5. S. P. Morgan, *Bell Labs. Record* **51**, 194 (1973).
6. J. L. Flanagan, *Proc. IEEE* **64**, 405 (1976).
7. See, for example, K. B. McAfee and R. Gnanadesikan, *A.I.Ch.E. Air Pollution Symp. Series*, in press.
8. D. Edelson, *J. Comp. Phys.* **11**, 455 (1973).
9. W. S. Cleveland, T. E. Graedel, B. Kleiner, J. L. Warner, *Science* **186**, 1037 (1974).
10. R. N. Shepard, A. K. Romney, S. B. Nerlove, *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, (Seminar Press, New York, 1972), vols. 1 and 2.
11. U. Neisser, *Cognitive Psychology* (Appleton-Century-Crofts, New York, 1967).
12. A. Newell and H. A. Simon, *Human Problem Solving* (Prentice-Hall, Englewood Cliffs, N.J., 1972).
13. S. Sternberg, *American Scientist* **57**, 421 (1969).
14. D. E. Knuth, *Science* **194**, 1235 (1976).
15. B. W. Kernighan, *Software—Pract. Exper.* **5**, 395 (1975).
16. B. G. Ryder, *ibid.* **4**, 359 (1974).
17. D. M. Ritchie and K. Thompson, *Commun. ACM* **17**, 365 (1974).

Tutored Videotape Instruction: A New Use of Electronics Media in Education

J. F. Gibbons, W. R. Kincheloe, K. S. Down

In the early 1920's, shortly after radio broadcasting was proved to be economically feasible, Robert Hutchins is said to have predicted that this new technology would undoubtedly have a dramatic impact on education. Subsequent events have shown that his assessment of the educational potential of radio was probably correct but, for a variety of reasons, the potential did not materialize. In the early 1950's instructional television was introduced with a similar fanfare. However, with a few notable exceptions, its potential also failed to materialize. It seems that more recent innovations such as computer-aided instruction and satellite-based educational delivery may come to a similar fate. Why is it that these technological aids to education seldom seem to live up to their potential?

There is of course a different set of reasons in each case, though inconstant financing and the competition with commercial interests are surely among the most pervasive. However, important as these factors are, there seems to be a still more basic problem. The proponents of media-based education describe this problem as a failure of the educational establishment to involve itself seriously with instructional technology. As a result, they say, the changes in the design of the educational system that must be made before instructional technologies can be used effectively have not been forthcoming. This is a valid criticism. However, the educational establishment makes a counterargument that is also true: The devices of instructional technology are too inflexible. Effective classroom teachers regularly capitalize on unexpected, unplanned opportunities for the achievement of specific goals. As Jackson (*1*) says, "Stray thoughts, sudden insights, meandering digressions and other unpredicted events constantly ruffle the smoothness of the instructional dialogue. In most classrooms, as every teacher knows, the path of educational progress could be more easily traced by a butterfly than by a bullet." Jackson concludes from this that education is best served by tools that can be readily adapted to a wide variety of educational tasks with a minimum of advance planning. Compared to most electronics media, blackboards and books provide at low cost an impressive degree of flexibility. Furthermore, after honest efforts to use electronics media over an extended period of time, many teachers have been unable to see a clear improvement in learning. Hence, electronics media are generally judged by teachers to be inappropriate educational tools for most circumstances.

Tutored Videotape Instruction

If we accept these criticisms as valid, we are led to seek out ways of using those media that will have the desired flexibility without requiring a major change in teaching styles, and to apply them to situations where the changes in the educational system that are necessary to accommodate them can be easily

Dr. Gibbons is professor and Dr. Kincheloe is adjunct professor of electrical engineering, and Mr. Down is assistant dean of engineering at Stanford University, Stanford, California 94305.

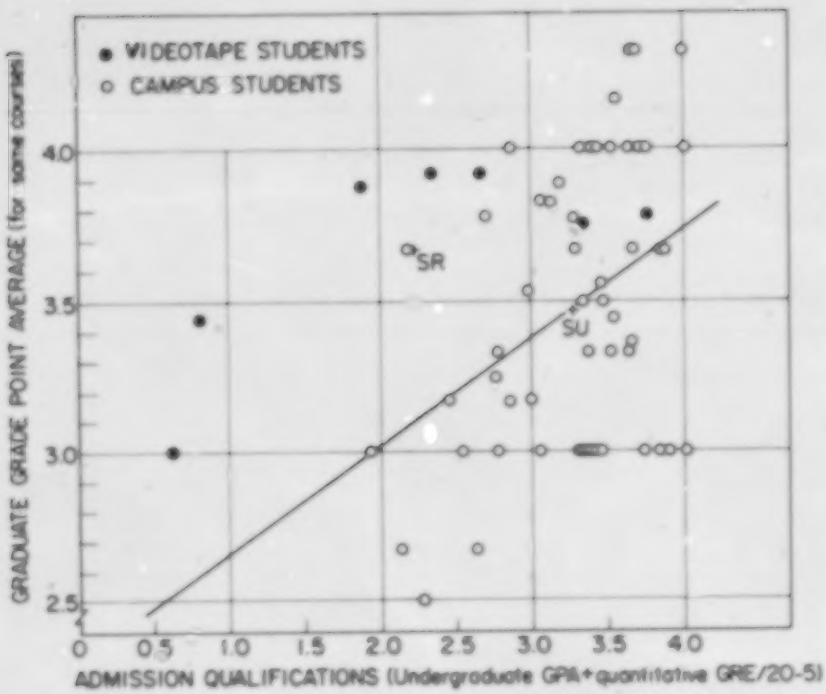


Fig. 1. Results of tutored videotape instruction experiment (October 1973 to March 1974).

made. In this article we describe a new technique for using videotaped educational materials that seems to meet these criteria. The method makes use of unrehearsed, unedited videotapes of regular classroom courses, which can be produced at very low cost. The videotapes are used for the instruction of small groups of students (typically 3 to 10) who are assisted by paraprofessional tutors as they watch the tape. For this reason, the method is called Tutored Videotape Instruction (TVI).

The TVI method was originally developed to provide course work in science and engineering to off-campus graduate students enrolled at Stanford University. It is based on the common-sense notion that students can learn more from a lecture if they are free to interrupt it at places where they need more discussion or explanation of a point or concept. Experience over the past 3 years shows that students learn best when the videotaped lectures are stopped frequently (for example, every 5 to 10 minutes, for periods of 3 to 5 minutes). Interactions of such frequency and duration are, of course, impracticable in a conventional classroom situation. Also, TVI is most effective when a small group of students and the tutor watch the tape together. It is less effective for a single student, with or without a tutor.

The TVI technique responds to the educational needs of the students by combining the positive features of lectures with those of small group discussions. The lectures provide for depth and continuity in the subject matter, while the tutorial discussions afford a means of making the lectures respond to individual needs and differences. Students watching the videotaped lectures feel free to ask questions of both the tutor and each other, and to make spontaneous comments about points of interest. In addition, the TVI students hear the comments that were made and the questions that were asked in the classroom at the time the videotape was recorded, and thus profit from those exchanges as well. In effect, the TVI format permits the students and the tutor to manage the lecture themselves and thereby create an intellectually stimulating environment which enhances learning and creates a positive attitude toward the subject.

Although the program to be described relies heavily on the use of video technology, that is probably not its most distinctive feature. Rather, the technology permits us to concentrate attention on some of the most important parameters in the learning process and to begin to understand their effects more precisely. What has been found seems to be applicable in a wide variety of circumstances. It extends well beyond university lecture courses and is in some ways independent of whether or not technological aids are used in the instruction process.

The program was designed to utilize insofar as possible the experience which has been gained over the past two decades concerning the question of how technology can be used most successfully in education. An attempt to summarize this experience led one of us to develop a set of guidelines that could assist educators in planning for the appropriate use of technology in education (2). Because those guidelines form the basis of the program to be discussed, it is useful to begin with a brief discussion of them.

Guidelines for the Application of Technology to Education

Current interest in the use of technological aids to education (particularly television) stems from the late 1940's and early 1950's when, following the widespread development of commercial television, the Ford Foundation and later the federal government funded a variety of experiments to study the possible impact of television on the educational process. These experiments included a complete range of educational institutions from elementary school to university, and took as their major emphasis the question of whether students could learn as well from television as from conventional classroom teaching. Surveying the results of 421 experiments carried out from 1950 to 1960, Chu and Schramm (3) concluded that, across all grade levels from kindergarten through a baccalaureate degree, and for essentially all types of subject matter, students could learn as well from television as from conventional instruction.

Viewed in its most positive light, this result suggests that access to education can be expanded substantially by communications technology, and indeed this policy has been urged by educational study groups and implemented by both private and public educational institutions for some time, and at substantial expense. However, the educational effectiveness of classroom instruction is itself a matter of increasing concern, and in this light a finding of "no significant difference in effect between television instruction and the regular classroom" does not provide a forceful stimulus to promote the application of communications technology to educational problems.

This raises the question of whether educational technology has ever estab-

lished a clear-cut superiority over in-class instruction. From a comparison of some of the most successful ventures in educational technology (such as Sesame Street, Chicago City TV College, the Open University in Britain, the Australian School of the Air, and the Bavarian Telecollege) with some of the least successful ventures, we attempted to abstract a set of guidelines that would describe the conditions that seem most likely to ensure a successful application of technology to education. These guidelines, while they may need further refinement and modification, do provide a useful starting point for designing new programs.

The guidelines are as follows:

1) The educational program should be planned for a specific audience.

2) Specific educational objectives that are relevant to the needs and interests of the audience should be clearly defined.

3) Technologies should be chosen in terms of the topic to be presented. Frequently, different technologies are used to present different parts of a course, the choice being determined from a consideration of which technology is most effective for the material being presented. It is desirable (though also expensive) to use both knowledge and media specialists to prepare and produce the programs.

4) Educators who have a clear interest in learning and using the instructional characteristics of various media should be selected and trained.

5) Clear and careful provision should be made for personal interaction, especially among the students.

6) Evaluation and feedback over a period of months or years should be used to monitor the educational effectiveness of the program, and the instructional materials and methods should be changed accordingly.

From the standpoint of educational effectiveness, the guideline that is perhaps most frequently overlooked is the one relating to personal interaction, especially where the use of television is concerned. Television (including lectures that are videotaped for subsequent playback) is most frequently used as a direct substitute for live lectures in large classes or as a means for individuals to view lectures or programs at sites that are remote from the point of origin. These applications are developed primarily to solve problems of cost and access, the assumption being made that courses delivered by television will be as effective as conventional education. However, the experiments that led to this conclusion tested only factual information gained over a relatively brief period. Furthermore, these experiments identified some characteristic weaknesses of television as an educational medium. In particular, as noted by Schramm (4), television does not stop to answer questions; it does not readily permit classroom discussion; it is an inefficient medium for conducting drill; it does not adjust very well to individual differences; and it tends to encourage a passive form of learning.

These are very serious weaknesses. They can be reduced to some degree by using radio talkback to connect remote sites to the television classroom (5), or by establishing a regular telephone contact between students and faculty, as has been done by the Chicago TV College and others. However, these techniques for promoting student-faculty interaction do not really remedy the major weaknesses described above. It is critically important for students to be able to stop a lecture or a program when they have questions, and it is highly desirable for it to remain stopped long enough for the question to be clearly answered. It is also highly desirable that the answers to the most important questions be developed as a result of active discussion within a small group of students. Finally, it is important that questions and discussions be used to determine background deficiencies of individual students, so that remedial action can be taken.

The major weaknesses of live television (including videotapes that are simply replayed from start to finish) are related to the fact that it cannot provide the quality of personal interaction that is available in a good classroom. The TVI intellectual community, consisting of a small group of students, an on-site paraprofessional tutor, and a course faculty member, was invented in an attempt to circumvent these deficiencies. What we seek to do is to provide small groups of students with the high-quality personal interaction that they need in order to learn effectively. What we have found is that TVI can provide a means of achieving this goal.

Experiments with TVI

We have carried out a number of experiments to evaluate the educational effectiveness of the TVI format. All the experiments to be discussed have been concerned with the delivery of engineering and science courses where the objectivity of the material makes it relatively easy to measure the effectiveness of the method. We believe the method can be extended to a much larger range of subjects and audiences, but there are insufficient data available at present to permit us to define this range of applications with any accuracy. Our description of experiments is therefore intended primarily to provide some general ideas about how the TVI format can be set up and evaluated.

The TVI technique was originally conceived as a means of providing courses to off-campus engineering students employed at a Hewlett-Packard (HP) plant in Santa Rosa, California, about 100 miles north of Stanford. The HP management wished to provide the same educational opportunities to these students as were available to HP employees (and those of other companies) at plants near Stanford where engineering courses given on campus are televised to plants within a 50-mile radius of the university by the Stanford Instructional Television Fixed Service (ITFS) network. The ITFS students also have an audio talkback connection that permits them to ask questions of the instructor while the class is in progress.

Since the courses of interest were already being televised, it was a simple matter to videotape the classes live with no further production requirements. It should be noted that an operating TV system is not a prerequisite for the TVI technique, though it was very helpful in this case because it both facilitated the initial experiments and made possible a thorough comparison between the regular TV students and the videotape students.

It was intended from the outset that satisfactory performance in the courses would result in credit toward a master's degree in electrical engineering from Stanford. Therefore, the Santa Rosa students were required to complete the same homework as students on campus, all their papers being graded by the same teaching assistant who graded the papers of the on-campus and regular TV students, and the off-campus students were required to come to the campus to take the same examinations as the on-campus students.

Program design for the Santa Rosa experiment. The main features of the program developed for the Santa Rosa experiment can be described in terms of the guidelines listed above.

1) The audience consists of industrially based students carrying full-time job responsibilities. In some cases the students were studying for a master's

degree before they were moved to Santa Rosa and had academic qualifications that were essentially identical to those of on-campus students, but students at Santa Rosa whose academic credentials were known to be inadequate for admission to the Stanford graduate school were also included.

2) In defining the educational objectives, we knew that the students were interested in graduate training leading to a master's degree. A clearly important factor in their success was that their employer shared this objective, in both its financial and academic aspects.

3) The technology used, half-inch reel-to-reel videotapes, was chosen to permit the clear reconstruction of a given TV frame at the remote sites. The videotapes were of live classes made by a trained student production staff as the class was being conducted. The classroom modifications that are necessary for this purpose are described by Pettit and Grace (5). Basically, the studio classrooms are organized to interfere as little as possible with the teaching styles and preferences of the instructors.

The videotapes are mailed to Santa Rosa along with class notes, homework assignments, and other materials that are handed out to the on-campus students. The tapes and homework are returned to Stanford in approximately 1 week. When necessary, the TVI tutors telephone the on-campus faculty after the videotapes have been watched to discuss problems and obtain supplementary material.

4) The educational staff whose courses are to be televised are given a brief training session to acquaint them with the capabilities of the television network and to offer suggestions concerning how these capabilities can be used effectively (how to organize blackboard space, use of a desk pad as an alternative to the blackboard, preparation of demonstrations, and so on).

In addition, Stanford staff members who are responsible for the TVI program visit each site to choose the tutors (from among the company's staff) and to instruct them in the use of the videotape as an educational medium. An alternative that is now being explored is to use as tutors staff members from local educational institutions who have indicated an interest in the program.

In most instances, the tutors are practicing engineers without prior experience in teaching. They are chosen primarily on the basis of two criteria: (i) a sensitivity to students and an ability to draw them into a fruitful discussion of issues, and (ii) a personal interest in reviewing the subject to be presented. Other criteria, such as recent exposure to the course with evidence of high-quality performance, have been found to be less important.

The tutor's main functions are: (i) to initiate and encourage stopping the videotape playback for the immediate resolution of problems; (ii) to answer, if possible, questions that cannot be resolved by the class; and (iii) to obtain answers and supplementary material from the on-campus instructor if necessary. Tutors are also encouraged to visit the on-campus faculty once or twice a quarter to become familiar with the course syllabus and to discuss any recurring problems that their students have. The tutor is not responsible for grading homework or assigning course grades.

5) Students and tutors are urged to stop the videotape whenever they have problems or questions or whenever some particularly important concept has been presented. Certain obvious cues are frequently used by the tutor to initiate these discussions. For example, each time a substantive question arises in the on-campus class, the tape is stopped and the TVI class attempts to generate the answer before the taped lecture proceeds.

In addition to extensive group interaction during the lectures, students are encouraged to discuss problems with each other and with the tutor outside the viewing period. Since the tutors and students are known to each other through their common employment, there is ample opportunity for them to do so.

6) Careful records are maintained of TVI student performance on both homework and examinations. Their performance is regularly compared with that of regular students in the campus classroom and local off-campus students receiving the same courses by way of ITFS.

The homework and test performance of the TVI students is analyzed by the Stanford TVI staff and the results are discussed with the tutors by telephone. Corrective measures are suggested when necessary. In addition, course evaluation questionnaires are given to the TVI students in order to assess their attitudes and reactions to the program.

Initial experimental results. After the first two quarters of operation, the TVI experimental program was evaluated by comparing the course performance of the TVI students to that of both the on-campus students and the HP students taking the same courses by live television. One result of this evaluation is shown in Fig. 1, which is drawn from data gathered during the autumn and winter quarters of the 1973–74 school year. This figure shows the grade point aver-

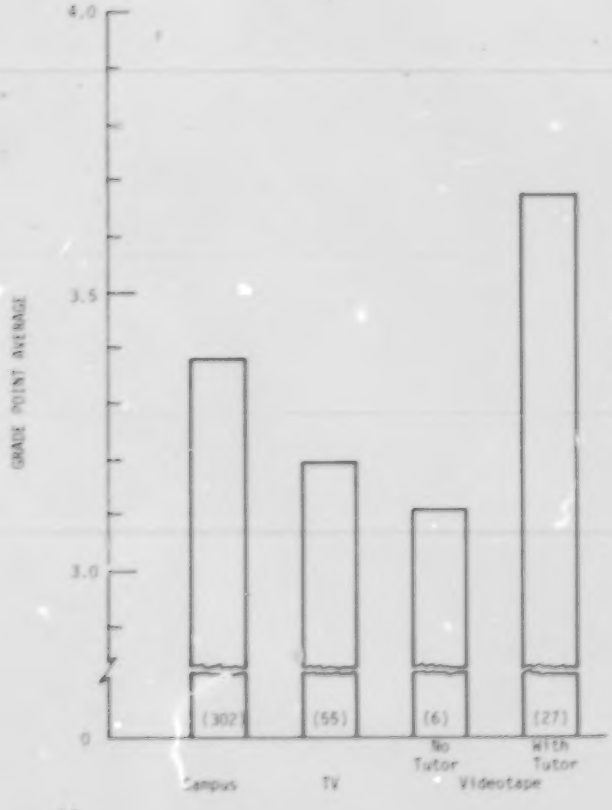


Fig. 2. Results of tutored videotape instruction experiment (October 1973 to March 1974).

ages of the TVI students for all courses taken under the program plotted against their admission qualifications, the latter being a composite of corrected undergraduate grade point averages (GPA's) and the quantitative component of the graduate record examination score. (The correction to the undergraduate GPA was based on prior history of performance of students from the same school in doing graduate work at Stanford.) Also shown in Fig. 1 are similar data for on-campus graduate students taking the same courses as the TVI students. The regression line shows the best linear least-mean-square fit to the on-campus student data, from which the average expected performance may be obtained for a student with given qualifications.

It is apparent from Fig. 1 that as a group the TVI students out-performed their on-campus counterparts; that is, the average GPA of the TVI students (represented by the point labeled SR in Fig. 1) is higher than that of the on-campus students (labeled SU in Fig. 1), even though the average of the admission qualification scores for the TVI students is substantially lower than that of the on-campus students. To make a more refined observation, it is convenient to divide the TVI students into two groups: those with admission qualification scores ≥ 2, who could have been admitted to the university as regular graduate students, and those with admission qualification scores substantially < 2 who would not have been so admitted. In terms of these subgroups, the data show that:

1) The subgroup that would have been granted admission did extremely well; in fact, their performance was essentially independent of their actual admission qualification scores.

2) The two students with the lowest qualification scores did quite acceptable work (B or better) even though they could not have been admitted to the Stanford Graduate School based solely on their admissions data. However, on the basis of their continued record of acceptable TVI performance, these students were subsequently admitted to the university and have now completed the master's degree program with creditable performance.

In Fig. 2, the grades of the TVI students are compared with those of the on-campus students and those of the students receiving the same courses by ITFS. The 302 on-campus students achieved a GPA of 3.38 out of a possible 4.00, which is typical of graduate electrical engineering students at the master's level. The students participating in the same courses by ITFS with audio talkback capability to the classroom had a GPA of 3.19, still quite acceptable but nonetheless below the on-campus students by almost a fifth of a grade point. The TVI students taking the same courses achieved a GPA of 3.67, almost half a grade point above the industrial student having comparable job situations and academic background. This result is made even more remarkable when we recall that several of the TVI students had marginal academic qualifications that would have made their admission to the Stanford graduate program very unlikely. It is also interesting to note that the performance of students at Santa Rosa in the videotape courses without local tutors was substantially below that of all the other groups, although the data in this case are very limited.

The total number of courses taken by students studying from videotape without a tutor is so small that the result has no statistical significance. However, a similar result has been obtained recently by Anderson (6) in a much larger experiment. In particular he finds that both student satisfaction and course performance tend to decrease as the delivery method is changed from on-campus lecture to live TV to nontutored videotape.

It is sometimes argued that the industrial experience of the TVI students accounts for their outstanding performance in these courses; but the TVI students are drawn from the same population as the students studying by ITFS whom they out-performed, so industrial experience or motivation cannot account for these results. Furthermore, other experiments show that on-campus TVI students also out-perform on-campus students who attend the regular lectures.

It is also natural to question the degree to which the results may be due to a Hawthorne effect. However, the Santa Rosa TVI students have continued their superior performance for 3 years.

Continuation of the TVI Industrial Program

Because of the initially encouraging results with the Santa Rosa industrial TVI students, the university gave permission to continue the program and to incorporate TVI courses into an accredited graduate degree program. New students at the Santa Rosa plant of Hewlett Packard entered the program in both 1974 and 1975; and the program was further extended to include Hewlett-Packard plants in San Diego and Boise and the Sandia facility in Albuquerque.

At the end of the 1975 to 1976 school year, after 3 years of operation, a total of 1803 quarter-units (approximately 600 courses) of graded course work, plus a considerable number of ungraded seminars, have been completed by 82 TVI students, with 65 different tutors having been involved in the program. The overall GPA of the TVI students for this work is 3.37. If we exclude from this group students whose academic qualifications were below those normally required for Stanford admission, the GPA rises to 3.59 compared to an on-campus average of 3.43 for the same courses. In other words, the TVI students who are qualified for regular admission (48 students) have continued to out-perform the on-campus students; and the poorly qualified TVI students have continued to do very acceptable work.

On the basis of these results it seems reasonable to conclude that, for science and engineering courses, the TVI format is at least as good as the other methods of delivery with which it has been compared. We believe the method can be used successfully for other types of courses, though we have no data to support this hypothesis. We have, however, used the technique in two on-campus experiments to determine whether full-time graduate students could also benefit from TVI.

On-Campus TVI Experiments

The first of these experiments was performed over two successive quarters in a graduate electrical engineering course that was large enough that several TVI groups, varying in size and with different tutors, could be formed for comparison with the large live lecture class. The course was taught by a faculty member who had already established a reputation as a particularly effective instructor for both regular classes and off-campus TVI groups. Three TVI groups were formed, two of these being led by the same tutor. The tutors were chosen from a set of students who had performed equally well in the course during the previous year. Both tutors were very much interested in the subject and were also interested in teaching as a career. However, their possible teaching styles were known from previous experience to be somewhat different. Where one tutor tended to answer questions directly when the tape was stopped, the other tended to encourage his group to find the answer. The first tutor was given a group of six students (labeled group 3 in the following figures). The second was given two

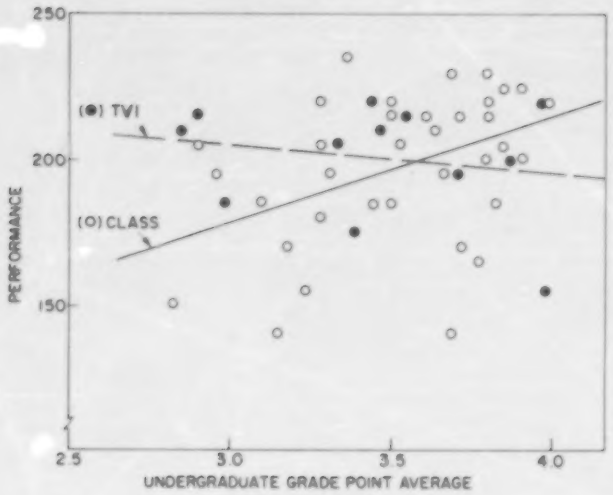


Fig. 3. Classroom and tutored videotape instruction student performance. Foreign students and honors cooperative students were eliminated from both the TVI groups and the in-class group for this comparison because they formed very different proportions of the groups.

groups, one with 12 members (identified as group 1) and one with six (group 2).

The separate TVI groups were too small for statistically significant differences to be manifested in their educational effectiveness. However, as illustrated by Fig. 3, the combined data show that the TVI method was at least as good as live instruction. Furthermore, for students with lower admissions qualifications, the TVI method of teaching appeared to be more effective than the regular classes. In fact the regression line for TVI students in Fig. 3 suggests that the course performance is essentially independent of the standard measures of academic ability. Note that this finding agrees with the results given in Fig. 1 for the subgroup of TVI students whose admission qualification scores were sufficient for them to be admitted to the university as regular graduate students.

Student opinions of the TVI format were also collected at the end of each quarter and showed a generally enthusiastic response to the method. Some characteristic features of their responses are shown in Table 1. With respect to the effect of group size, a comparison of groups 1 and 2 (same tutor) shows that the smaller group was generally more enthusiastic about the TVI experience than the larger group. This same general result has been repeated in our industrial TVI groups, and a consideration of all of our present data on group size leads us to the tentative conclusion that the method works best when the group size is between 3 and 10.

Table 1 also contains some specific information about the influence of tutoring styles that is consistent with our general observations. The students studying with the tutor who tended to answer questions directly (group 3) were less pleased than students in a group of equal size with a tutor who tended to draw them into the discussion when the tape was stopped.

The second set of experiments, performed during the winter quarter of 1976, was conducted with the second half of an introductory graduate engineering-economics systems course in which the regular course professor was away from campus. In the absence of a suitable lecturing replacement, videotapes of his course from the previous year were shown to the regular class. A professor from the general subject area viewed the tapes with the class and answered questions primarily at the end of the lecture. For technical and scheduling reasons, the tapes shown to the large class could not be stopped during the lecture for thorough discussion.

Table 1. Student reactions to tutored videotape instruction on campus. Question A was, "How would you rate your overall educational experience with TVI as compared to a large live lecture class?" Question B was, "How do you feel about taking another course in this manner?" To obtain averages and standard deviations a preference scale running from +1 (definitely superior or highly favored) to −1 (definitely inferior or definitely do not prefer) was employed.

Rating	Groups 1 (N = 12)	2 (N = 6)	3 (N = 6)
Question A			
Definitely superior	4	3	
Superior	6	3	1
About equal	2		4
Inferior			
Definitely inferior			
$\bar{X}$	0.58	0.75	0.1
σ	0.36	0.27	0.22
Question B			
Highly favor	7	6	3
Favor	3		
Neutral	2		2
Prefer not			
Definitely not			
$\bar{X}$	0.71	1.0	0.6
σ	0.4	0.0	0.55

With respect to course grades, no statistically significant difference was found in the performance of the TVI and control groups, partly because of the small size of the sample groups, although again the data were consistent with the conclusion that the TVI students did at least as well as the regular class, and that the students of lower ability benefited more by the TVI educational method. In fact the comparison was frustrated by the fact that all low-ability students in the control group dropped the course.

However, a pronounced difference was observed in student attitudes toward the use of a videotape with and without the combined tutor and tape-stopping features. These results are shown in Table 2, where the responses to several questions are recorded for the three groups of students (low ability, high ability, and control). Each question was answered on a preference scale which ran from +1 (strongly agree) to −1 (strongly disagree). Note that the high-ability TVI group was on the average highly enthusiastic on all counts, though its course performance was almost identical to what it had been during the previous quarter in the live lecture class. In fact, the high-ability TVI group stopped the tape on an average of every 5 minutes.

Observations

The TVI technique was invented in an effort to provide the benefits of both lectures and small group discussions to off-campus engineering students. Experience gained from 3 years of operation of the program suggests that the TVI technique is at least as effective as either classroom instruction or live TV with audio talkback capability, for both on-campus and off-campus students. However, our data do not yet permit a rigorous statistical test of this conclusion to be made. We are also unable to generalize to subject areas other than engineering and science, though we believe the general principles of the TVI format will

apply to a wide range of subjects and audiences.

To assist in considering how the method might be applied to situations other than those described above, we have enumerated the factors which we believe to be critical to the effectiveness of the TVI format:

1) The attitude, personality, and instructional style of the tutor are very important. The tutor should be interested in helping the students in his group. He should attend all or nearly all the videotape sessions. His competence is important, but it is better that he not be so overqualified that he becomes bored or impatient with a lack of understanding in the students. Compensation of the tutors is important for a continuing program.

2) Group size is also very important. If there are fewer than three students opportunity for effective interaction is lacking and the method tends to be expensive. Group size greater than 8 to 10 tends to inhibit discussion and reduce the frequency with which the tape is stopped. A group size of 3 to 8 seems optimum, although this can vary with student personalities and acquaintance with each other.

3) Depending on the maturity of the student, commitment to a degree program or similar educational objective seems to be important for sustaining interest and motivation. Certainly for most students completion of graded problems and examinations results in a more productive educational experience.

4) Active classroom participation in the live class is desirable. For the subjects and audiences served to date, unrehearsed, unedited videotapes of classroom lectures may be used and, in fact, may have more "presence" and be more interesting to watch than tightly scripted, professionally produced lectures.

5) It is important that the instructor be well organized, knowledgeable in his subject, and free of annoying mannerisms. The charisma of a good instructor is emphasized on the videotape.

6) For students employed in industry, attitudes of management play a very important role in the success of a continuing program. Job pressures that create long hours and interfere with family life markedly increase the difficulty of pursuing an educational program.

7) Continued management and evaluation of a TVI program needs to be the concern and principal responsibility of a designated person who should provide liaison between the academic institution and the TVI students. Many details require timely attention that would otherwise not be given by either the instructor or the company.

None of these factors pose unmanageable requirements. With attention to group size, good tutoring, quality of recorded educational material, adequate handling of supplementary materials, and grading of problems and examinations, the TVI methodology can provide an excellent educational experience and opportunities for needed education in otherwise difficult or impossible situations.

Given that it can accomplish important educational objectives, it seems worthwhile to summarize the principal advantages TVI has over both in-class instruction and live television:

1) It is (or can be) educationally more effective than either classroom or televised instruction, especially for marginally qualified students. This is primarily because it permits students to ask more questions and, through organized discussion, to find more answers for themselves than they could in either of the other formats.

2) TVI is cheaper than either classroom or televised instruction from the point of view of marginal costs. It does not require new teachers, new educational plant, or expensive broadcast facilities. Our operating experience shows that TVI can be provided at a marginal cost of $2.20 per student per lecture for groups of 10 students. The marginal cost of instruction by "real time" television on the Stanford ITFS system is $2.43 per student per lecture.

3) It makes good use of teaching resources by using faculty for course preparation and paraprofessional tutors for discussion of the lectures. For this latter function it draws on the substantial (if latent) interest in tutoring that exists in a large segment of the population.

4) It allows instruction to take place at the convenience of the students. It is not bound by either the academic calendar or a broadcast schedule.

These advantages are substantial, and while over-generalizing is dangerous, we nonetheless feel confident that TVI can be successfully extended to large-scale applications, at least for courses in science and engineering. We believe the method can also be successfully extended to other subject areas.

Table 2. Campus student attitudes toward tutored videotape instruction. The abbreviations are: SA, strongly agree; A, agree; N, neutral; D, disagree; SD, strongly disagree. The attitudes examined were: 1, feel more favorable toward this method of instruction at end of course than when course began; 2, prefer this method of instruction to lectures in large classes (50 or more students); 3, learned more from this course than from other courses due to the instructional method; 4, asked more questions in this class than in other classes; 5, learned more from questions and comments of other students than in other courses; 6, felt more free to ask questions and express myself than in other courses; 7, would recommend this course to a friend exactly as I took it. A numerical performance scale was established for quantification of the data: SA = 1, A = 0.5, N = 0, D = −0.5, SD = −1.0; $\bar{X}$ and σ are means and standard deviation, respectively.

Group	SA	A	N	D	SD	$\bar{X}$	σ
			Question 1				
Low	2	3	2	0	0	0.5	0.41
High	1	4	1	0	0	0.5	0.32
Control	0	7	6	6	3	−0.1	0.54
			Question 2				
Low	4	2	1	0	0	0.71	0.39
High	5	1	0	0	0	0.92	0.20
Control	1	6	5	5	3	−0.075	0.59
			Question 3				
Low	2	2	2	1	0	0.36	0.56
High	4	1	0	1	0	0.67	0.61
Control	1	3	6	10	2	−0.2	0.5
			Question 4				
Low	0	5	1	1	0	0.29	0.39
High	3	3	0	0	0	0.75	0.27
Control	0	0	10	7	4	−0.36	0.39
			Question 5				
Low	2	4	0	1	0	0.5	0.5
High	2	3	1	0	0	0.58	0.38
Control	0	0	5	9	8	−0.57	0.39
			Question 6				
Low	4	1	1	1	0	0.57	0.61
High	4	1	1	0	0	0.75	0.42
Control	0	0	8	7	7	−0.48	0.42
			Question 7				
Low	3	3	0	0	1	0.5	0.71
High	1	2	2	1	0	0.13	0.74
Control	1	4	7	6	4	−0.18	0.57

References and Notes

1. P. W. Jackson, *The Teacher and The Machine, Horace Mann Lecture, 1967* (Univ. of Pittsburgh Press, Pittsburgh, 1968).
2. J. F. Gibbons, "Federal initiatives for bringing communications technology to the service of education," position paper for the Newman Task Force on Higher Education and the President's Science Advisory Committee. A partial summary of the paper is published in *Aspen Notebook on Cable and Continuing Education*, R. Adler and W. S. Baer, Eds. (Praeger, New York, 1973), pp. 127–132.
3. G. C. Chu and W. Schramm, "Learning from television: What the research says" (Educational Resources Information Center Report No. Ed-014-900, Stanford University, Stanford, 1967).
4. W. Schramm, "What we know about learning from television," (Educational Resources Information Center Report No. ED-002-561, Stanford University, Stanford, 1971).
5. J. M. Pettit and D. J. Grace, *IEEE Spectrum*, May 1970.
6. R. M. Anderson, Jr., personal communication.

7. We thank J. G. Linvill for encouragement and support, J. Lindsay and R. Lotta for assistance with the statistics, and several of our colleagues for helpful reviews of the manuscript. The National Institute of Education provided support for the statistical analysis presented and also commissioned a paper from which this article was abstracted.

Educational Uses of the PLATO Computer System

The PLATO system is used for instruction, scientific research, and communications.

Stanley G. Smith and Bruce Arne Sherwood

The PLATO (*1*) computer-based educational system has been specifically designed to provide interactive, self-paced instruction to large numbers of students (*2*). Lesson material is displayed on a screen 22 centimeters square and may consist of text, drawings, graphs, and color photographs. Students interact with the material through a special keyset that closely resembles a typewriter keyboard, and they receive essentially instantaneous reinforcement of correct work and assistance where they are having difficulty. Students can work at their convenience in classrooms such as the one shown in Fig. 1.

The users of PLATO range from grade school students learning reading and math to graduate students in the medical sciences. The system now has 950 terminals located in universities, colleges, community colleges, public schools, military training schools, and commercial organizations (*3*). The users have access to more than 3500 hours of instructional material in more than 100 subject areas (*4*). We will mainly describe one area of PLATO use—that of university science education and research.

Examples of PLATO Lessons

The character of PLATO lesson material varies greatly since the computer system does not impose a pedagogical structure on the authors of the materials. Some appreciation of the breadth of approaches used may be gained by reviewing brief segments of a few programs in chemistry (*5*) and physics (*6*). The examples below are illustrated with photographs of the student's plasma-panel screen (*7*). Unfortunately, however, these static photographs do not fully convey the dynamic nature of the interactively changing displays seen by the student.

A physics lesson on oscillations contains features common to many expository science lessons. The student is given a table of contents for the lesson so that he may choose to study any section or even go directly to a final quiz that will test his understanding of the material. The intent of the lesson is to take students from no knowledge of oscillating systems up to a point where their understanding allows them to solve typical homework problems.

The first part of the lesson deals with an oscillator consisting of a block sliding up and down two smooth inclines, a system that the student analyzes by simple kinematics formulas. In order to help the student understand the system, the computer shows an animation of the motion and allows him to experiment with values of the initial displacement (Fig. 2).

After the student has experimented with the system until he thinks he understands it, the program tests his knowledge by asking a number of questions about the system. Help is provided where necessary, but where help is required the student must answer the same question later with different numerical factors. Such checkup sections follow each expository section of the lesson.

The final quiz is constructed from questions asked in the earlier sections with randomly chosen numerical factors. No help is provided. The lesson distinguishes between incorrect numerical results and typing errors, such as unbalanced parentheses. In this lesson, if the student misses more than two out of six questions, he must take the whole quiz again. He can, of course, review sections of the lesson if he wishes.

Students also have the opportunity to study systems of their own design and to program the computer with a special language so that they may obtain immediate graphical results. Figure 3 illustrates an example of large-amplitude pendulum motion in which the student has used s for angle, g for gravitational acceleration, l for length, v for angular velocity, t for time, and d for a short time interval. The graphs that have been produced for the student correspond to running the program with a starting angle of 45 degrees and with a starting angle of 179 degrees.

Since much of science is based on the results of experiments, it seems important to have students learn to design experiments and interpret the experimental data. However, many of the key experiments in the development of important concepts cannot be carried out by large numbers of students because of the lack of adequate equipment, time, and experimental technique. The use of computer simulation can serve to provide some experience with the concepts. For example, a comparison of the sensitivity of the rates of ethanolysis of *n*-butyl bromide and *t*-butyl bromide as a function of ethoxide concentration serves to clarify some aspects of the concepts of unimolecular and bimolecular reactions.

Dr. Smith is professor of chemistry at the University of Illinois, Urbana 61801. Dr. Sherwood is associate professor of physics and assistant director of the Computer-Based Eduction Research Laboratory at the University of Illinois, Urbana.

Students can quickly discover what happens through simulated experiments in which the computer plots the percent reaction as a function of time as would be observed under the experimental conditions they suggest. Since the student is free to explore the relation between experimental conditions and reaction rate, it is important to have the program ask questions and, if necessary, suggest additional experiments to assure that a suitable set of experiments has been done. In this case, after the experimental facts have been established and a suitable interpretation has been developed, the system provides a visual picture of the transformations involved by means of an animation that shows the sequence of bond making and breaking which occurs.

In more advanced lessons, students are given a problem that can be solved by conducting some simulated experiments on the computer. The student is expected to design the experiment, select the compounds and reaction conditions, and then collect the experimental data, do the mathematical analysis, and outline the conclusions that can be derived from his experiments. This approach to teaching is possible because the computer can rapidly calculate the outcome of experiments of a given type from algorithms which describe the results of actual experiments that may be beyond the experimental skills of the students and the available laboratory facilities.

It is also necessary for students to gain experience in dealing with problems for which there are many possible solutions. The synthesis of organic compounds is an example of this type of situation, in which there are many viable routes from the starting materials to the designated product. This is illustrated in Fig. 4 where the student is given the task of converting a given starting material into a designated product molecule. He proceeds by suggesting a reagent for each step in the transformation. Since there are many possible paths between the starting material and the product, the computer is programmed (8) to carry out the reaction suggested by the student and compare the structure of the product with that of the desired material. If they are the same, the synthesis is judged completed. If not, the reagent for the next step is requested. This approach does not impose a specific solution to the problem on the student but recognizes a wide range of acceptable routes.

One way to provide practice problems for organic reactions is illustrated in Fig. 5 where the student has been given 16 compounds and ten reagents and is asked to find at least 20 different ways of interconverting pairs of the compounds with the use of the reagents shown.

Fig. 1. One of the PLATO classrooms at the University of Illinois. This classroom in the Foreign Language Building contains 80 terminals.

To cause the interaction, the student simply points to a compound and then to the reactant or product side of the reaction arrow to indicate his choice. The computer senses which compound the student is pointing at by means of a matrix of infrared light-emitting diodes and sensors that lie in a 16 by 16 array around the edge of the display screen. The use of such a "touch" sensitive display makes it easy and quick for the student to specify the reactants and products. Errors are corrected by having the computer show either the correct product for a given reactant and reagent or the required reagent for a reactant product pair.

Additional flexibility in the use of the computer as an aid to teaching comes from the ability to have the computer control the projection of color photographs on the plasma panel. This is illustrated in Fig. 6, which has been abstracted from a lesson on the use of an analytical balance. In this example the student must identify the function of the knob on the side of the instrument.

Structure of Lessons

The successful application of program segments such as those described above to a real teaching situation involves the integration of practice problems and associated help sequences with the development of the necessary theoretical framework to assist the students in understanding the material. The ability of the PLATO system to support a one-to-one dialogue with the student offers the possibility of presenting the material in unique new ways that make the student an active participant in an effective learning situation. Well-designed lesson material tends to be highly interactive and requires frequent inputs from each student in the form of answers to questions, predictions of the outcome of some experiment, parameters to be used in simulated experiments, and interpretation of a set of data or facts. In addition, since the understanding of the subject matter that a student has before starting a given program varies enormously, it is desirable to structure the program to accommodate students who just need a brief review as well as those who are learning the material for the first time. One simple way to provide students with flexibility in the way they study and use a given program is to provide an index to the lesson which allows easy access to any section.

Since students can proceed at their own rate, time spent in the lesson, unlike a lecture, automatically adjusts to the needs of the student. In fact, students proceed at greatly different rates when given the opportunity. For many lessons the time required for students to complete the lesson often varies (9) by a factor of 3 to 7.

Although lesson material can make some adjustments to meet the needs of individual students, it is important to develop criteria and data that indicate how the difficulty of the programs matches the abilities of the students. One criterion that has been used is the percentage of the questions posed within the program which students answer correctly on their first try. The assumption in this approach is that if nearly all student responses in instructional material are correct, then the programs are not adequately challenging them, while a very small

percentage correct suggests that the lesson is unduly discouraging. The plot of percentage OK on the first try as a function of the number of students, shown in Fig. 7, suggests that for most of the students the level of the material was adequately adjusted for the class (*10*).

In addition to allowing multiple entry points and the ability to review as frequently as desired, lessons should tend to adjust to each student within each section. For example, help should be provided either when requested by the student (there is a key on the keyset labeled HELP) or when it is clear that he is having difficulty. The number of problems presented can be easily adjusted to the student by such simple means as requiring that he get two right in a row of a certain type or, perhaps, simply by returning a given problem to the list of those that need to be worked if the student needed assistance in working it. Data from lessons are used to check that the lesson is adjusting properly to the students' needs.

Computer-Managed Instruction

A complete lesson on PLATO has many of the characteristics of a chapter in a textbook. Like chapters in a book, such

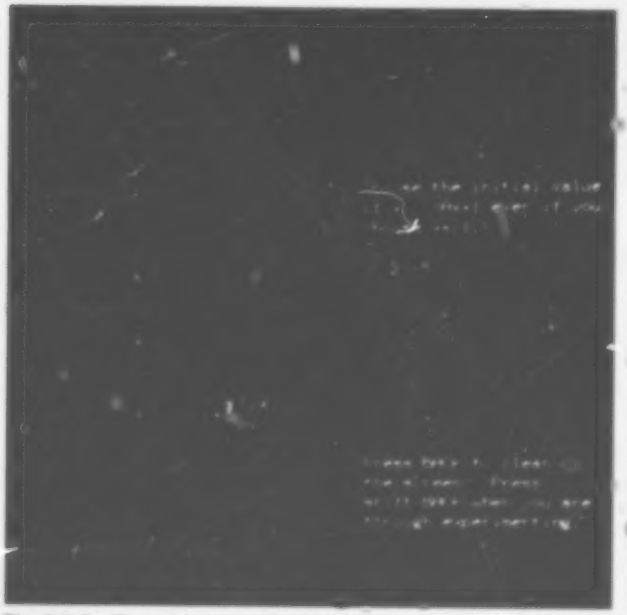


Fig. 2 (left). The student can choose the amplitude of this nonlinear oscillator and study the effect of amplitude on frequency. Fig. 3 (right). The student has written a numerical integration program to study the motion of a pendulum with large-amplitude swings. The two graphs correspond to amplitudes of 45 degrees and 179 degrees.

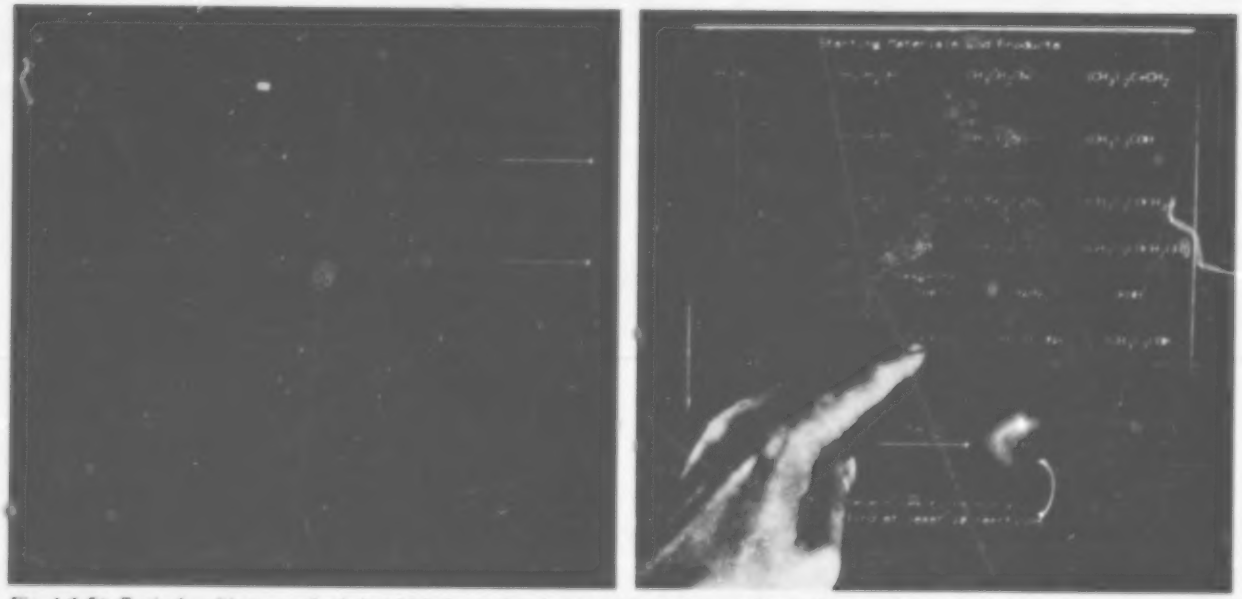


Fig. 4 (left). Typical multistep synthesis in which the student has typed in the reagent for each step as the computer draws the structure of the reaction product. Fig. 5 (right). The student can specify the starting material, reagent, and product by simply touching the compound on the screen. A 16 by 16 array of infrared light-emitting diodes and sensors determines where the student is pointing.

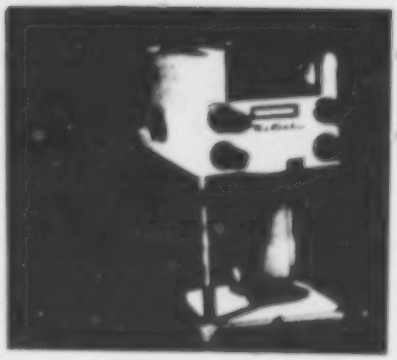


Fig. 6 (right). The student must give the function of the knob on the side of the analytical balance. The computer controls the projection of color photographs from a microfiche onto the back of the plasma panel.

lessons need to be assembled in a form that is easy for students to use. All of the lessons associated with a particular course can be made available from an index. Then, when a student signs on to the PLATO system with his name and the name of the course, he may choose topics or lessons to study. He makes his selection from a list of descriptive titles, much as chapters in a book are selected from the table of contents. (As mentioned above, many lessons have, in addition, a table of contents for the subsections of the lesson.)

If there are a large number of lessons associated with a particular course, it may be desirable to provide the student with some guidance in the selection of lessons that are appropriate to the course at that time. The PLATO system makes this possible by a course management scheme that allows an instructor to set up an index of lessons by simply selecting the lessons from a catalog of lessons displayed on the screen. Many such indexes may be set up for a course. For example, all of the lessons associated with a given topic or concept may be placed on one index. The instructor then specifies the criteria for allowing students to move from one index of lessons to another. For example, the student might be required to complete three of four lessons before moving ahead to the next topic. Or, this criterion may be modified so that at a given date the next set of lessons is made available even if the specified number of earlier lessons has not been completed, so that a student who has gotten behind at one point in the course can keep up with the new material. If on-line quizzes or exams are included in a given module or set of lessons, a satisfactory score can be included in the criteria specified for completion before new topics are presented.

The relations among the student, instructor, and the lesson author are diagramed in Fig. 8. This scheme, which is available to all instructors, provides guidance to the student on current work, makes it easy to review earlier lessons, and allows students to work ahead of the rest of the class. The result is an efficient and effective integration of the techniques of computer-based teaching and computer-managed instruction. While this generally available management scheme is used by many instructors, it is also possible to create other management structures to meet special requirements.

Integration of PLATO Activities

PLATO has been integrated into the structure of courses in several ways. For example, in a classical mechanics course (*11*), computer-based tutoring made it possible to drop one of the two weekly lectures. The remaining lecture is used chiefly for demonstrations rather than for basic instruction. The discussion period is spent in the physics PLATO classroom (which has 30 terminals), where students work individually but can get help from the instructor. Students spend additional study time at a terminal on a nonscheduled basis. Thirty terminals used 60 hours per week, providing 4 hours of contact to each student, can serve 450 students. Large numbers of terminals are required to make an impact on instruction.

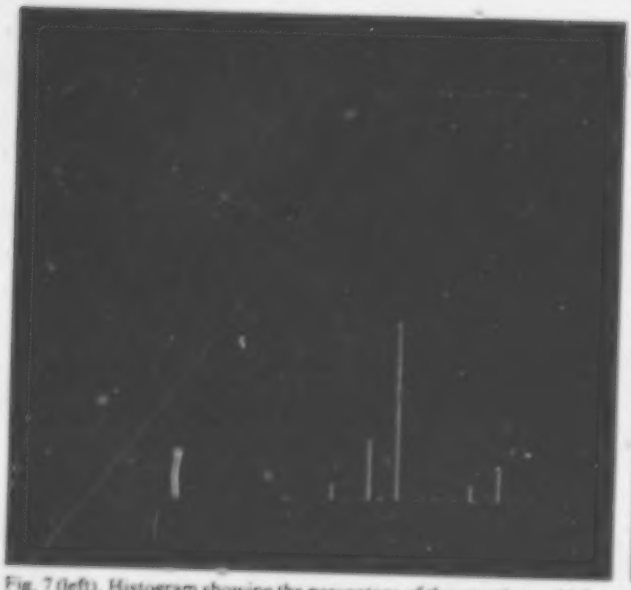
There are three main components of the PLATO aspects of the course: instructional lessons, homework, and an on-line gradebook. Students are assigned instructional lessons to study, and part of their grade is based on how many of these lessons they complete. Homework is graded by PLATO rather than by the instructors. The student is given printed homework problems that he is encouraged to work at home. When the student is ready, he goes to a terminal to enter his results. If the problem involves numerical quantities, each student has different numbers. A convenient calculator is always available. The student obtains a numerical score on the homework; homework scores form another basis for the course grade.

Both lesson completion data and homework scores flow automatically into an on-line gradebook. Instructors also enter other grades into this gradebook, such as exam scores and lab report scores. Each student can look at his own scores and can see his position in a graphical display of distributions throughout the course to compare how he is doing with respect to the rest of the students. Each instructor can look at and change the scores of his own students and can see scores for his section marked on the course distribution graph. The overall director of the course can look at the status of any section. One important benefit of this machinery is that an instructor can plan class activities on the basis of up-to-date information on how far students have gotten in their studies. Instructors also get rapid and accurate indications of students who are falling behind.

Just as the same textbook is used in association with many different kinds of courses, PLATO materials are integrated in various ways with other activities. The same PLATO physics materials used at the University of Illinois in the rather structured scheme described above are used at Carnegie-Mellon University in a self-paced course where PLATO instructional lessons are simply another resource for student study. In addition, the modularity of PLATO lessons makes them easy to use with diverse textbooks.

PLATO has been integrated into instruction in many other ways. For example, in language courses (including French, Spanish, German, Russian, Hebrew, Latin, and Esperanto), PLATO is used heavily to drill the student on vocabulary and grammar and to give practice in translating sentences from one language to the other. A two-semester course designed to teach students how to read Russian consists of a standard textbook plus PLATO lessons for each chapter of the textbook (*12*). This reading course is an alternative track to the general language course for undergraduates and graduate students who wish only to read Russian. Optional laboratory drills for the standard beginning Russian course are also on PLATO.

The function of the PLATO lessons is not only to give practice but also to test, with instant feedback, whether the student understands the concepts presented in the textbook. The materials are also used for review, either within the course, or by persons who want to refresh their knowledge of Russian. There is normally a class discussion before and after each textbook lesson, occurring every one-and-a-half to two weeks. The student spends the bulk of his study time at the PLATO terminal. In all areas, PLATO lessons are usually integrated with additional classroom activities. However, in such environments as continuing or adult education where, because of constraints



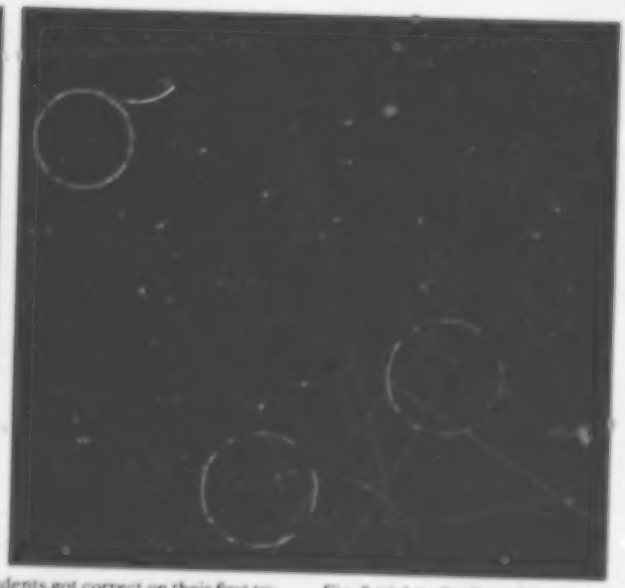


Fig. 7 (left). Histogram showing the percentage of the questions which students got correct on their first try. Fig. 8 (right). Outline of the PLATO computer-based education system. Students who have been put on the roster have access to lesson material selected by the instructor from the catalog of lessons. Students have access only to their own grades in the gradebook, which can automatically collect scores from lessons. Data related to the lesson performance are stored for review by the author of the programs.

of time or distance, the student may not be able to participate in scheduled classroom activities, a course consisting only of PLATO materials is a viable alternative.

Another important integration of PLATO into courses is illustrated by the use in chemistry of simulated laboratory experiments as a means of better preparing the student for a real laboratory experiment. Merely simulating an experiment would be inappropriate if the manual techniques themselves and not just the intellectual content are important. A PLATO lesson on the theory of titration is followed by a PLATO simulation of an actual titration, in which the student must specify all the steps, including controlling the flow from the buret, and then analyze the data. Errors made in doing the simulated experiment lead to the same problems that would be observed in the laboratory. These preparatory activities are then completed by going into the laboratory and actually performing the experiment, at a time when the student has a thorough understanding of the content of the experiment and can concentrate on the practical complications that arise. PLATO lessons may also be used to extend laboratory experience by simulated experiments which, because of limitations of time, facilities, and experimental skills, students are not able to do in the laboratory. One such example is shown in Fig. 9, which is taken from a lesson on fractional distillation.

There are ambitious curriculum development projects for reading (*13*) and mathematics (*14*) in the elementary school. Both of these projects constitute highly integrated packages of instructional materials, with automatic routing from one activity to another (based on performance) and extensive reporting to the teachers.

Writing Lesson Material in TUTOR

Lesson material for use on the PLATO system is written in the TUTOR programming language (*15*) which has been designed to facilitate the development of interactive instructional programs on a graphic computer system. As a further aid to authors, who should be primarily concerned with the problems of teaching and not with learning about computers, professional consultants can be reached easily through a PLATO terminal. The consultant, who can be at another terminal anywhere on the system, can see the program on the author's screen and type and receive comments at the bottom of the screen. In addition, a very complete random access, cross-indexed description of TUTOR and its use is available by simply typing one of more than 2000 key words that describe the type of thing on which information is desired (*16*). The key words do not have to be spelled correctly, yet information is delivered to the requester almost instantaneously. More than 1600 such requests per day are answered by the system. A single key press returns the lesson author to the point in his work where he requested information. Another crucial feature is that an author can construct a display consisting of text and line drawings on the screen, and PLATO will generate automatically the TUTOR program corresponding to that display. The extensive on-line help for authors of lesson material makes the PLATO system a programming aid to authors as well as a learning aid to students. One result is that it takes less time to create quality lesson materials than has been typical with earlier systems.

To demonstrate some of the features of the TUTOR programming language, below is an example of the complete computer program to provide a chemistry student with a simple problem:

```
unit    reagent
at      710
write   Indicate the reagent required for
        this reaction.
at      1318
write   C6H6 → C6H5Br
arrow   1225
answer  (Fe, FeBr3, ferric*bromide)
        <and> (Br2, bromine)
wrong   (Br2, bromine)
write   You also need some Fe.
no
writec  ntries-2,
        This is an electrophilic bromina-
        tion.,
        Try Br2 and Fe.
```

The program says that "at" line 7 of the 32 lines on the screen, spaced over 10 of the 64 horizontal spaces, the computer should "write" "Indicate the reagent required for this reaction." Then at line 13, space 18, the chemical reaction is drawn

on the screen. The command "arrow" indicates the location on the screen where the student's answer is to appear. With the "answer" command are listed the common reagents that the students would be expected to know. Words enclosed in parentheses are regarded as synonyms while words enclosed in[printed brackets < > may be present in the answer but are not required to be there. The phrase "ferric*bromide" is treated as a single word. The "wrong" is an example of a typical anticipated wrong answer for which the response given in the next "write" statement will be given. The "no" followed by "writec" instructs the computer to judge the answer so if the indicated reagents are not given and to write a message contingent ("writec") on the number of times the student has tried ("ntries") to answer the question. Here, the author has decided to first provide a clue, then a suggested correct response.

This example involved judging a student's sentence. Other TUTOR judging techniques used heavily in science lessons include judging algebraic and numeric responses, with optional dimensional analysis of answers involving scientific units.

Programs dealing with chemistry and physics often involve displaying graphs for students to interpret or plotting functions or data supplied by the student. The short program given below will plot the amplitude response of a driven oscillator as a function of the driving frequency. The student can choose the time constant associated with the damping of the system. Figure 10 shows the result of this program where the student has superimposed several graphs. The basic display components of the program, such as the positioning on the screen of the axes and the graph titles, were created automatically by PLATO; the author merely positioned these elements on the screen and PLATO created the corresponding program statements.

```
define    z = v1, c = v2
unit      resonant
origin    75,200
axes      360,250
scalex    2
scaley    8
labelx    1..2
labely    1..5
delta     .02
at        2252
write     ω/ω0
size      1
rotate    90
at        1505
write     Amplitude
size      0
at        2316
write     Choose a value for the time constant (in units of 1/ω0):
arrow     2720
store     c
ok
funct     1/[(c² − 1)² + (c × z)²], z
```

For accuracy in plotting, the alternative fine grid coordinate system is used in which ($x = 0, y = 0$) is the lower left corner of the screen, and (511,511) is the upper right corner of the screen. The command "origin" specifies that the origin of the plot will be at $x = 75$ and $y = 200$. The "axes" command defines the length of the x and y axes. The axes are scaled to the units of the problem by means of the "scalex" and "scaley" commands. The "labelx" and "labely" indicate how the axes should be labeled, and the "delta" command establishes an increment to be used in plotting the function. The "size" and "rotate" commands permit writing a rotated label ("size 0" restores normal writing options). The "store" command evaluates mathematical expressions typed by the student and stores the result in the variable "c." The "funct" command instructs the computer to plot the stated function of z ($z = \omega/\omega_0$) for values of z within the range of the x-axis.

PLATO as a Communication Medium

The highly interactive character of the PLATO system has promoted the development of efficient communication systems between the users. These communication systems take several forms. There is a general note file where authors may ask questions or make suggestions for changes to the system. Similar forums for exchanging notes of general interest among smaller groups of users, such as all authors in a given subject area, are also widely used. Perhaps the most important communication system between users comes from the ability to send a personal note, which can be several pages long, to another individual on the PLATO system. The "letter" is delivered instantaneously to the addressee at whatever terminal he is using or uses next. Although the message is delivered immedi-

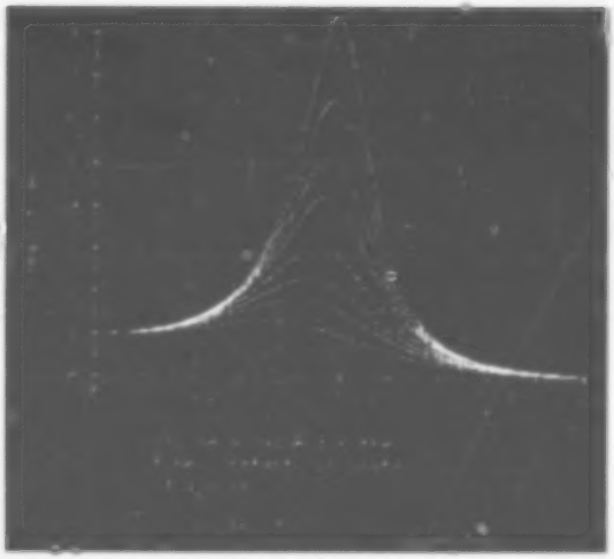


Fig. 9 (left). In a lesson on fractional distillation, the computer has reminded the student that he needs to add a boiling chip before proceeding with the distillation. Fig. 10 (right). The student can specify a damping factor and see a plot of the frequency response of a driven oscillator. The program required to produce this display is described in the text.

ately, it need not be answered until it is convenient. It is not at all uncommon for several exchanges of "letters" to occur across the country within a single day. On a typical day, more than 1000 such personal notes are written on the PLATO system.

If immediate discussion is desired, Smith can page Sherwood by name. If Sherwood is at a terminal (and has specified that he will accept calls), he receives a message at the bottom of the screen informing him that Smith (of the chemistry group) wants to talk to him. Sherwood has the option of accepting or rejecting the call. Smith does not have to know where Sherwood is located because the computer system can find him. If the call is accepted, the two authors can converse by typing at the bottom of each other's screens. To ensure privacy, neither author can see the rest of the other's screen unless it is explicitly offered, in which case Smith can watch Sherwood's screen (or vice versa). Being able to see what is on the other person's screen is extremely useful for discussing programming and lesson design problems. Both authors can continue to type comments at the bottom of the screen.

Computational Uses and On-Line Data Collection

The characteristics of the PLATO system which make it particularly useful for the development of instructional material also make the system a versatile computing tool in research laboratories (*17*). The TUTOR language makes it easy to write programs to do calculations to analyze the data, and special graphing commands coupled with the full graphic display facilitate accurate construction of graphs. These may be reduced to hard copy with a special terminal (made by Varian) which contains an electrostatic copier.

The introduction of experimental data from laboratory equipment is accomplished through a connector on the terminal which allows two-way communication between the computer system and peripheral equipment. Currently the PLATO system can accept an average of about three keys per second per terminal (students enter about 0.5 key per second), so the rate of transfer of data from peripheral equipment is limited to about 25 bits per second. This means that, for experiments which gather large amounts of data rapidly, a buffer is necessary between the experiment and the terminal to avoid exceeding the bandwidth of the computer system. The data can, of course, be stored permanently on the system, recalled, processed, and displayed in graphical form on the terminal.

The speed and ease of use gives the impression of dealing with an on-line computer which, however, has the extraordinary calculating and graphical capability of a very large system. It may be possible that a PLATO-like system will meet the needs of many laboratories for an on-line computer system.

It is also possible to run standard computational jobs on the computer at the same time that interactive PLATO services are being provided. On the Urbana system this additional computational capability has been used only for internal purposes, but it is planned in the long run to make such services more generally available.

Development and Distribution of Lessons

The cost of the development of curriculum material for the PLATO system is mainly the time required by the teacher to create the programs. A college teacher experienced in the PLATO system will spend 10 to 50 hours at a terminal while creating an hour of instruction, with some additional planning time spent away from the terminal. That is, the time or money costs of producing a PLATO lesson, which is one "chapter" of a PLATO course, are probably similar to the costs of producing one chapter of a textbook (*18*). Generally speaking, curriculum costs have been much higher on other computer-based systems, but PLATO offers significantly improved aids to authors.

There has been much study of the severe problems associated with distributing educational computer programs from one institution to another. Problems include lack of compatibility from one computer system to another and lack of adequate communications among institutions. The PLATO system has finessed the problem by tying large numbers of terminals to one center, thus providing compatibility automatically and making possible extensive electronic communications. The intention is to tie all PLATO systems together to form a national grid of regional networks, thus ensuring the same level of compatibility and communications.

Accounting procedures are now in effect which log usage of lessons, both to provide authors with information on which institutions are using a lesson and to provide a basis for paying royalties to authors. If PLATO services are sufficiently widespread, a few cents per hour of use of a lesson could provide significant income to authors. This would stimulate production and make it self-sustaining without artificial subsidies. There would also be competition among similar materials, which is extremely important. Only a small fraction of published college textbooks are considered adequate. It is likely that good computer-based materials will also be rare since similar creative processes are involved.

Electronic publishing does differ in some interesting ways from paper publishing. Often an instructor would like to pick chapters out of various books to assemble in his own preferred sequence. This is difficult with paper but easy with electronically published lessons.

Value of PLATO-Based Instruction

The value to an institution of the broad range of services offered by a PLATO system cannot be easily quantified. It is particularly difficult to measure the value of the direct instruction provided by PLATO since there is an enormous number of variables involved. It is also difficult to make a direct comparison with other teaching methods, partly because little quantitative data exist of the effectiveness of traditional teaching methods and, also, because some components of a PLATO-based course do not exist in traditional courses (one-to-one dialogues and simulations of physical phenomena, among others). Nevertheless, it is useful to summarize some basic observations about PLATO-based science instruction.

First, students do as well or better in PLATO-based science courses compared with similar courses that do not use PLATO. They turn in more work: nearly all students complete all assigned instructional lessons, and nearly all students turn in all assigned homework exercises. In a physics course on classical mechanics, 400 students taught by traditional methods and 200 PLATO students took the same 3-hour final exam. The median score of the PLATO students was slightly higher but not statistically significant, and the distributions of scores were very similar. Comparable results were obtained in the previous semester, with half as many students involved.

Second, students like PLATO-based science instruction, as reported consistently in formal questionnaires and in informal comments. Since many students have multiple contacts with PLATO-based courses, and since such courses typically

involve several hours of PLATO use per week, it seems likely that student approval of science instruction with PLATO is not due simply to its novelty. Students report that they believe that PLATO instruction is helpful in learning the material.

Third, faculty like PLATO-based science instruction. Discounting the enthusiastic authors of the materials, other instructors who have had the opportunity to use PLATO materials in their courses say that they and their students find the materials useful. This is true not only at the University of Illinois but also at the many other universities and community colleges around the country where these materials are in regular use.

It must be emphasized that these encouraging results have been obtained during a period when the techniques for using this new medium have only begun to be exploited. The interactive PLATO medium is quite different from previous passive media such as movies, television, and lectures, and it may take many years to develop this new medium to its full potential.

Costs of PLATO Services

Because basic research and development costs are intertwined with the cost of providing direct instruction on the experimental PLATO system, it is difficult to estimate the cost of providing classroom instruction alone. However, the major components that contribute to the cost of the operation of the system can be identified and used as a guide to potential costs in other environments. These are the terminals (*19*), the central computer facility (*20*), maintenance, and salaries for the computer center staff. In addition, terminals that are located far from the central computer have extra costs for telephone connections, while local terminals use a microwave system (*21*). The actual cost of delivering instruction with such a system will depend on the current price of these components, the fraction of the facilities which is devoted to other activities such as batch computation, and the organization operating the system.

Although the PLATO system requires a large computer, the cost of such an installation is offset by heavy usage of more than a million terminal hours per year (*22*). Last year the average usage was 1300 hours per terminal, averaged over all terminals in many different environments. Taken as a whole, the 300 terminals at the University of Illinois averaged about 40 hours of use per week, but the two semesters last a total of only 30 weeks, and summer use by students is light. Some terminals at the University have utilization rates as high as 2000 hours per year.

The costs of establishing additional PLATO systems are also affected by rapid changes in technology. The Control Data Corporation Cyber computer in use by PLATO is basically unchanged from its introduction 10 years ago so that more modern technology could result in substantial cost reductions. The terminal, which is also a major factor in the cost, represents new technology in small-scale production, so there exists the potential for cost reductions in this area.

How PLATO Works

The terminal now being used with the PLATO system uses a plasma display panel (*7*) that has a grid of 512 by 512 electrodes at whose intersections a neon discharge can be ignited or extinguished. Text is written on these screens at the rate of 180 characters per second. Graphs and line drawings are displayed at the rate of 60 connected lines per second. Since any part of the display can be changed independently of the rest of the screen, it is possible to use animations to illustrate important concepts in lesson material. The plasma panel has inherent memory and does not require repetitive replotting of the screen as is necessary with normal cathode-ray tubes. The flat plasma panel also serves as a screen for the rear projection of color photographs that can be superimposed on text and drawings generated by the computer.

The fact that the PLATO system handles hundreds of graphical display terminals is the result of an unusual structure (*23*). In standard time-sharing systems, users' programs are swapped in and out of the computer to an electromechanical storage file (disk or drum). PLATO, on the other hand, uses Control Data Corporation's Extended Core Storage as the swapping medium. A unique operating system (which includes the TUTOR language) takes advantage of the much higher performance capabilities of this electronic memory, whose transfer rate is a hundred times faster than that of disks or drums, and whose access time is a thousand times shorter than the waiting time to retrieve information from disks or drums. This makes it possible to handle hundreds of graphical display terminals, with a response time of one-eighth second.

Summary

The PLATO computer-based teaching system provides individualized instruction to hundreds of students simultaneously. Lesson material can include natural language dialogue, graphics, numerical and algebraic responses, and color photographs. There now exist more than 3500 hours of tested materials in more than 100 subject areas. There is also a wide range of tools available for managing courses oriented to PLATO. Communications facilities include public forums, electronic mail, and on-line consulting. The system can be connected to laboratory equipment for on-line data collection and analysis.

There already exist three PLATO systems (*1*). It is expected that additional systems will be set up on a regional basis, with electronic intersystem connections to assure a high level of communication, including nationwide access to curriculum materials.

References and Notes

1. D. Alpert and D. L. Bitzer, *Science* **167**, 1582 (1970); A. L. Hammond, *ibid.* **176**, 1110 (1972); D. L. Bitzer, B. A. Sherwood, P. Tenczar, in *New Trends in the Utilization of Educational Technology for Science Education* (Unesco Press, Paris, 1974).
2. People who have been involved in the development of the "hardware" associated with the PLATO system include D. Bitzer, J. Stifle, F. Ebeling, M. Johnson, R. Johnson, F. Propst, D. Skaperdas, G. Slottow, and P. Tucker. A number of people have contributed to the development of the PLATO "software." Full-time people have included P. Tenczar, D. Andersen, R. Blomme, J. Carstedt, R. Chabay, C. Fugitt, D. Lee, R. Rader, M. Secrest, B. Sherwood, D. Shirer, and M. Walker. They have been assisted by W. Golden, J. Parry, D. Brown, B. Fortner, D. Frankel, S. Freyder, S. Gooch, D. Kopf, K. Mast, P. Mast, M. Midden, L. North, L. Steinberg, L. White, and D. Woolley.
3. There are additional PLATO systems in Minneapolis, Minn. (operated by Control Data Corporation), and in Tallahassee (operated by Florida State University).
4. E. R. Lyman, *PLATO Curricular Materials* (Computer-Based Education Research Laboratory, Urbana, Ill., 1975), report X-41, No. 3.
5. S. G. Smith, *J. Chem. Educ.* **47**, 608 (1970); *ibid.* **48**, 727 (1971); ———— and J. R. Ghesquiere, in *Computers in Chemistry and Instrumentation*, J. S. Mattson *et al.*, Eds. (Dekker, New York, 1974).
6. B. A. Sherwood, *Am. J. Phys.* **39**, 1199 (1971); ———, C. Bennett, J. Mitchell, C. Tenczar, in *Proceedings of the Conference on Computers in the Undergraduate Curriculum* (Dartmouth College, Hanover, N.H., 1971), p. 463; C. D. Bennett, in *Proceedings of the 1972 Conference on Computers in Undergraduate Curricula* (Southern Regional Education Board, Atlanta, 1972), p. 269.
7. J. Stifle, *Proc. Soc. Inf. Disp.* **13**, 35 (1972); R. L. Johnson *et al.*, *IEEE Trans. Electron Devices* **18**, 642 (1971). The plasma panel is thin, transparent, flat display device.
8. E. J. Corey and W. T. Wipke, *Science* **166**, 178 (1969).
9. R. Chabay, thesis, University of Illinois (1974); J. Ghesquiere, thesis, University of Illinois (1973).
10. S. G. Smith, J. R. Ghesquiere, R. A. Avner, *J. Chem. Educ.* **51**, 243 (1974).
11. People responsible for preparing the PLATO version of this course include C. Bennett, D. Kane, B. Sherwood, and J. Smith.
12. C. Curtin, D. Clayton, C. Finch, D. Moor, I. Woodruff, *Mod. Language J.* **56**, 354 (1972).

13. P. Oberlinto, *Educ. Technol.* **14**, 2 (1974); J. Risken and E. Webber, *ibid.*, p. 19.
14. R. B. Davis, in *Proceedings of the 1974 National Computer Conference* (American Federation of Information Processing Societies Press, Montvale, N.J., 1974), p. 169; S. Dugdale and D. Kibbey, *The Fractions Curriculum of the PLATO Elementary School Mathematics Project* (Computer-Based Education Research Laboratory, Urbana, Ill., 1975).
15. P. Tenczar, *Society for Information Display International Symposium '70* (1974); J. Ghesquiere, C. Davis, C. Thompson, *Introduction to TUTOR* (Computer-Based Education Research Laboratory, Urbana, Ill., 1974); B. A. Sherwood, *The TUTOR Language* (Computer-Based Education Research Laboratory, Urbana, Ill., 1974); P. Tenczar and W. M. Golden, *Spelling, Word and Concept Recognition* (Computer-Based Education Research Laboratory, Urbana, Ill., 1977), report X-35; E. Avner, *Summary of TUTOR Commands and System Variables* (Computer-Based Education Research Laboratory, Urbana, Ill., 1974).
16. This feature of the PLATO system is the responsibility of D. Chirdan, C. Davis, J. Ghesquiere, T. Gutnadius, J. Krautz, and J. Sherwood. Many people have made contributions to the documentation.
17. D. Walter and R. McKown, *Proceedings of a Symposium on Current Computer Technology* (MIT Press, Cambridge, Mass., in press).
18. George M. Grimes, *Cost of Initial Development of PLATO Instruction in Veterinary Medicine* (Computer-Based Education Research Laboratory, Urbana, Ill., 1975), report X-43.
19. Through a special purchase by the University of Illinois, terminals cost about $6000, including plasma panel, keyset, touch panel, and microfiche projector.
20. The computer used is a Control Data Corporation Cyber 73 with 2 million words of extended core storage.
21. The microwave distribution system costs about $2.50 per month per terminal.
22. R. A. Avner, personal communication.
23. B. Sherwood and J. Stifle, *The PLATO IV Communications System* (Computer-Based Education Research Laboratory, Urbana, Ill., 1975).
24. Many people have made significant contributions to the development of the PLATO system described here. None of the work would have been possible without the interest and help of D. L. Bitzer, director of the Computer-Based Education Research Laboratory at the University of Illinois. The work was supported by the University of Illinois, the National Science Foundation, the Advanced Research Projects Agency, and the Control Data Corporation.

Impact of the Electronics Revolution on Industrial Process Control

Acceptance of advanced digital communication and control systems in the industrial plant will accelerate.

Lawrence B. Evans

The first computer control system went on-line in an industrial plant in 1959. Since then, there have been remarkable advances in our ability to acquire, process, and transmit information electronically. Developments in the technology of digital computer hardware, software, basic sensors, and all forms of communication offer the potential for industrial process control systems that are highly automated and provide improved operating performance.

The Role of Process Control in Industrial Plants

The focus in this article is on the "process industries" as opposed to the manufacturing industries. Products of the process industries include chemicals, petroleum, metals, electric power, pulp and paper, food, cement, and textiles. Their plants manipulate the composition of materials by chemical reaction, purification, and blending of components to convert raw materials and energy into more valuable products. A modern chemical plant is shown in Fig. 1. The earliest applications of computer control were in the process industries, where instruments were available to monitor the continuous flow of a product and to send the data to the computer, which could then direct changes in the process by adjusting valves and switches.

Automation in the manufacturing industries—automobiles, appliances, electronics—is beyond the scope of this article. These industries manipulate the geometries of their raw materials so that discrete parts are assembled to form products. Computers are being used on the factory floor to run machine tools, track the contents of a warehouse, test products, and so forth, but the methods of measurement and control are basically different from those used in the process industries, and the problems of automation are greater (*1*).

Industrial processes are designed to operate in either a continuous or a batch mode. In the first, materials flow continuously through the plant from one processing unit to the next. At each stage, different operations are performed, such as heating, cooling, mixing, chemical reaction, distillation, drying, and pressurization. For most operations, there are optimum conditions of temperature, pressure, and residence time at each stage.

In the batch mode, the material being processed stays in one place (such as in a reaction vessel), and the process steps are carried out over time during the batch cycle. There is an optimum schedule of such factors as temperature and pressure. For large-scale production, engineers have traditionally tried to develop continuous processes, because they are easier to instrument and control, require less labor, and do not waste time in emptying, cleaning, and refilling vessels between batch cycles (*2*).

Flow rate is by far the most common variable manipulated, whether by adjusting a valve, turning a pump on or off, or by other means. The measured variables in addition to flow rate are normally temperature, pressure, chemical composition, and liquid level. A typical plant, such as one for manufacture of ethylene or ammonia, will have several hundred control valves and more than a thousand measured variables. Changes are made in the operation of the process on a time scale ranging from a few seconds to a few hours.

The elements of a control system are shown in Fig. 2. The important functions are measurement, control, actuation, and communication. Measurement refers to the sensing of variables such as flow rate, temperature, pressure, level, and chemical composition, and the transmission of the measurement to the controller. Control is the decision-making operation; it compares the measured state of the process with the desired conditions and decides how the variables should be manipulated. Actuation is the means by which the operating variables are manipulated; typical actuators are valves, rheostats, switches, and relays. Communication includes the display of information to the plant operators as well as the transmission of important variables to the plant management.

The author is professor of chemical engineering at Massachusetts Institute of Technology, Cambridge 02139.

80

The organization of a plant control system in a hierarchical structure is shown in Fig. 3. The lowest level is occupied by the control computer that regulates a single process unit to hold it to the desired operating conditions and move the unit to a safe condition in emergencies. The next step is a computer responsible for coordinating several units, for scheduling operations, and for optimizing the plant's performance. At the top level is the corporate control computer, which makes available to management current information about its manufacturing operations.

The operations at the lowest level (of regulating the process and watching for system failures) are the simplest, but they must be done frequently. The optimizing and scheduling tasks at the higher levels are much more complex but are done less frequently.

An advanced automation system is concerned not only with the physical control of the process but also with its management (the gathering and application of both process and business information for decision-making).

There are four basic objectives of plant operation that an advanced management and control system can help achieve. (i) To perform the primary mission of the plant. The plant must produce products of the required quantity and quality and be available and operating when needed. If it doesn't perform its primary mission, the plant is a failure even if succeeding objectives are met. (ii) To maintain safe operating conditions. The control system should constrain processing to safe operating conditions, such as allowable temperatures and pressures, and should be able to shut the system down safely in an emergency. (iii) To comply with government requirements. Society imposes restrictions on the operation of most industrial plants in the form of government regulations. These restrictions require the control of environmental emissions to acceptable levels, control over the exposure of workers to toxic materials, the keeping of records required by law, and so forth. (iv) To improve productivity. This can be achieved by increasing the yield of products, reducing requirements for raw materials, energy, and labor, and reducing capital investment. The greatest opportunities for advances in the application of modern control technology are in those situations that provide obvious benefits in meeting one or more of these objectives.

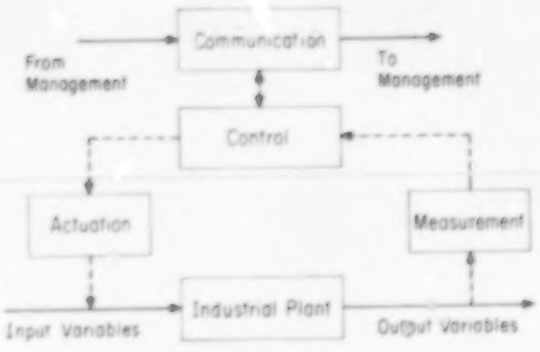
Fig. 1. Modern petrochemical plant. [Courtesy of the Foxboro Company]

Advances in Control Technology

There have been significant advances in control theory, in industrial applications of control, and in hardware and software available for control. Each of these areas will be reviewed.

Almost all of the early control systems, such as the flyball governor developed by James Watt in 1788 for controlling the speed of a steam engine, were developed empirically. Maxwell in 1868 (*1*) wrote the first theoretical paper on control with his analysis of the stability of the flyball governor. Since that time there have been rapid advances in the mathematical theory of control. These have led to improved understanding of the behavior of complex feedback control systems and to the development of mathematical and computational techniques for their design. Major developments in control theory are listed in Table 1. Advances in the application of control technology to industrial processes are summarized in Table 2.

When continuous processes began re-



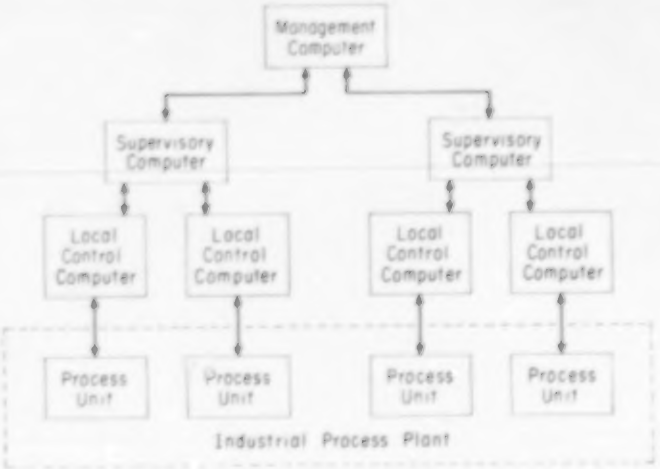


Fig. 2 (left). Schematic diagram of an industrial control system.
Fig. 3 (right). Hierarchical organization of an advanced control system.

placing batch stills in the oil industry in 1910, most processes were controlled manually with use of local temperature and pressure measurements. In the late 1920's the pneumatic controller was developed with proportional and integral (reset) modes that could be tuned in the field. A few years later the derivative mode was added to form the so-called PID (proportional-integral-derivative) three-term controller (*13*). This was a turning point in process instrumentation and led to a tremendous extension of the use of automatic control to processes previously difficult or impossible to control except manually.

Pneumatic transmitters were developed in the mid-1930's and led in the 1940's to centralized control rooms (Fig. 4) with large control panels. Electronic controllers became available in the 1950's. Since installation of the first computer-controlled system in 1959, there has been a steady growth in the number of computers installed in process plants.

An important historic observation is that significant developments took place essentially at plants. The development of the three-term pneumatic controller in essentially its modern configuration was completed in 1942 without the a priori support of any theory.

Among the recent developments in hardware and software for control are decreased cost of electronic hardware, improved communications capabilities, techniques for producing less expensive software of better quality, and the development of specialized sensors.

The cost of complex electronic circuitry continues to decrease exponentially (by a factor of about 1/2 each year) due to large-scale integration (LSI) semiconductor technology. Such technology has led to more powerful microprocessors with greater memory and speed and smaller cost per unit. The real cost of a system is in the hardware for communication between man and that system (displays, keys, typewriters) and this cost is a function of the way the system is packaged. Thus, automation functions and data processing become economic if they can be done blindly, without the need for human communication.

Improvements in cathode-ray tube (CRT) displays are continuing to permit displays of higher quality with greater flexibility at lower costs. Color displays are becoming much less expensive. The problem will be to determine how to present information to the operator so that it will be most useful.

Improved communications are making it possible to use systems in which elements of the control system are dispersed throughout the plant and communicate with each other through networks. New technologies for communication based on the use of coaxial cable, fiber optics, and other carriers are emerging. Communication over these networks is by a digital rather than an analog signal.

The major cost of a computer control system is in the engineering and programming of the system. A new technology for producing computer programs, known as "software engineering" is becoming established. Instead of programmers writing their programs in whatever way they think best, software is designed and produced by engineers who use proven techniques. This approach reduces software costs, errors, and implementation time while improving reliability and maintainability.

Analytical instruments are increasingly used to measure stream composition and condition at the output of a process for on-line control. Sensors are being developed to measure all sorts of specialized compositions, such as flue gas, sulfur in oil, and brightness. Frequently the availability of such a sensor is the key to developing an advanced control system for a particular process.

Trends in Process Plants

The use of advanced control technology has been influenced by developments in industrial processes and in the constraints imposed by society on their operation. The search in industry is for plants that produce more valuable products in greater quantity at lower cost. Plants are consequently becoming larger and more complex and must be operated to closer tolerances. To achieve economies of scale in the production of chemicals that are important commodities, such as ammonia and ethylene, plants of enormous size have been developed in which the processes are designed as a single train. The cost of a malfunction that stops production for any length of time is prohibitive.

The need to reduce pollution and to conserve energy has led to increased recycling and exchange of heat between process streams. Plants are becoming more tightly integrated so that changes in one part of the process have an effect in many other places. The control system must account for the resulting interaction.

Processes are being developed that involve more complex sequences of chemical operations requiring precise conditions of temperature and pressure. The increased use of catalytic processes requires controls to maintain conditions that will preclude catalyst degradation and poisoning.

The production of more sophisticated and valuable products requires closer control of product specifications. To manufacture Polaroid's new color film, nine layers of chemicals no more than a few 10,000ths of an inch thick are deposited on the film as it passes through a large coating machine. As the layers are applied, the sheet of film must be alternately dried, heated, and cooled under extremely precise computer control of temperature and humidity.

All these trends in major industrial

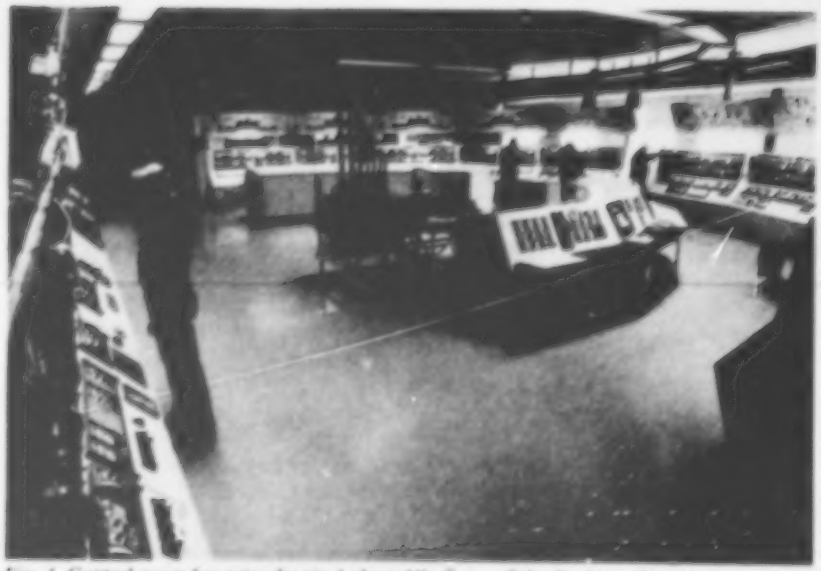


Fig. 4. Control room for petrochemical plant. [Courtesy of the Foxboro Company]

processes lead to requirements of more sophisticated control. Social and political considerations affect the way industrial plants are designed and operated. Important long-term forces are the needs for protection of the environment, conservation of energy and resources, safety, and good working conditions.

Because of the need to reduce emissions, it is now economical (in fact essential) to control the contents of waste streams. In the past, plant effluents have been the only streams not placed on automatic control.

The energy crisis has caused many companies to reevaluate their processes. New ground rules for designing plants favor tighter energy integration and the corresponding need for more sophisticated control.

Concern for safety works both for and against automation. Arguments are raised that it is not safe to run a plant without operators present and that an unattended plant would never get past a safety review board of the Office of Safety and Health Administration. However, sophisticated, properly designed systems can get people out of noisy, dirty environments, and this favors automation.

In some parts of the country, there are simply not enough skilled people available to serve as operators. People are less willing to devote their lives to the same employer and geographic region. Thus, industry can no longer depend on having an operator with years of experience available to assume responsibility. This encourages the use of more automated processes that are easier to operate.

Advanced automation affects job enrichment in two ways. On the one hand, it eliminates some drudgery, but, on the other, operators tend to get bored with computer systems. A highly automated system must be carefully designed if it is to enrich the job.

Industrial policies on the proper role of the operator in running plants vary, and there is corresponding variation in practice. There are two extreme approaches: (i) to automate the process as much as possible and use a relatively unskilled operator who could do little more than shut the plant down in case of an emergency; and (ii) to use a better trained, perhaps college educated, operator who would also serve as a manager. The ultimate direction industry will take is not clear at this time.

On balance, the social and political forces for environmental protection, energy conservation, safety, and better working conditions favor greater automation.

Rate of Acceptance

In recent years there have been some striking successes in the application of advanced digital communication and control systems in industrial plants. Computers have become commonplace in refineries, steel mills, chemical plants, and most process industries. They are no longer considered revolutionary or glamorous but have become an accepted part of the process control.

Each year the *Oil and Gas Journal* (*14*) publishes a list of the computer control systems in operation, being installed, or contracted to be installed worldwide in refineries, petrochemical plants, pipelines, and producing plants. There has been a steady increase with a growth rate of about 25 percent per year (Fig. 5). These figures include only a portion of the process industries, and many computer control projects are kept secret for proprietary reasons. But they indicate a trend. A recent estimate (*15*) states that today there are well over 200,000 digital computers worldwide in on-line industrial control applications (not just in the process industries) with the number increasing by 67 percent annually.

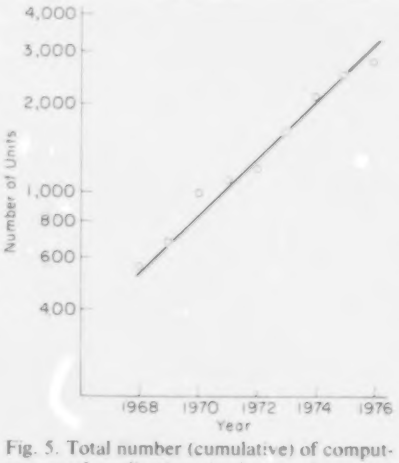


Fig. 5. Total number (cumulative) of computer control applications in the petroleum and petrochemical industry (*14*).

Despite the successes and the large number of process control computers that have been installed, there is still a great untapped potential for introduction of advanced computer-based automation systems into the industrial plant. Most observers in the industry agree that the acceptance of advanced technology is far less than the state of the art would permit.

In the late 1950's and early 1960's there was a burst of enthusiasm and a feeling that process control was about to enter a new era. Vannah and Slater (*16*) observed that "process control today is at the start of a new era—an era of analysis, introspection, and basic improvement." Yoxall (*17*) remarked that "We stand upon the threshold of a new decade and there is a general feeling that we shall see during these next ten years, significant developments in the specialized field of instrumentation." It was visualized that there would be computers connected to processes, that these processes would be optimized, that there would be improved techniques for measurement.

The new technology has not been accepted at anything approaching the rate forecast. McMillan (*18*) in 1966, writing a review of on-line computer control in the process industry, noted that "It is clear that the explosion never came and probably never will. Real progress has been

Table 1. Developments in control theory.

Date	Development
1868	Maxwell (*3*) read his paper, "On Governors," the first analytic study of the flyball governor.
1922	Minorsky (*4*) wrote the differential equations describing automatic ship-steering gear and used it to design a system on the *New Mexico*.
1932	Nyquist (*5*) presented his basic regeneration theory, which established him as the father of feedback control technology.
1934	Hazen (*6*) published the theory of servomechanisms, a term which he coined.
1936	Callender, Porter, and Hartree (*7*) published an early study of the transient performance of a valve-controlled process.
1938	Philbrick (*8*) developed the electronic analog computer at the Foxboro Company.
1953	Mason (*9*) introduced the signal-flow diagram techniques.
1956	Boltyanskii, Gamkrelidze, and Pontryagin (*10*) introduced generalized optimization synthesis techniques for dynamic systems.
1958	Kalman and Bertram (*11*) pointed out the utility of state-space representation of the dynamics of control systems.
1961	Kalman and Bucy (*12*) presented a new approach to filtering and prediction for linear systems.

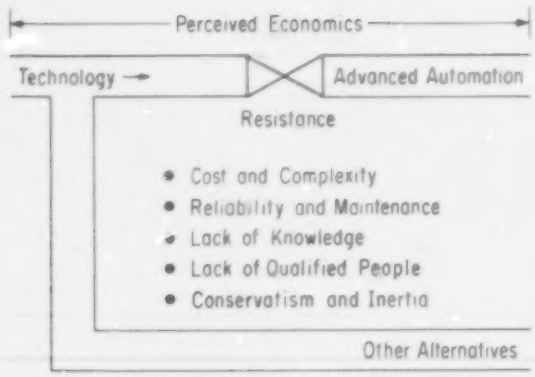


Fig. 6. Model of factors influencing acceptance of advanced automation.

made, but, as always, it continues to be hard-won."

While the practitioners in the field were lamenting the lack of use of advanced computer hardware, the control theorists were also complaining that modern control theory was not being used to its potential. A number of papers sought to explain the gap between theory and practice [for example (*19, 20*)].

Blocks to Acceptance of New Technology

A major stumbling block to the application of advanced control systems has been our lack of understanding of the processes to be controlled. To apply single-loop controllers, one needed to know only the direction of a change. (If the fuel supply to the boiler is increased, the steam pressure rises.) The application of advanced control, however, requires refined process knowledge and mathematical models.

Another block has been a shortage of trained people who both understand the industrial processes and also have knowledge of and sophistication in computer control.

A third limitation is the inertia and conservatism in the industry. The heavy process industries with large capital investments are hesitant to take technical risks because the penalty for failure is so large. Some companies had bad experiences with systems installed in the early days of computer control that failed to deliver as promised; it may take a generation to overcome this resistance.

Part of the resistance to advanced control systems is technological. It is economically difficult to provide systems that are reliable and maintainable. It may well be technically feasible to achieve the required degree of reliability, but for some applications such systems are so expensive they are no longer practical.

The lack of standardization in control systems and plant designs poses yet another problem. Because each control system is custom designed, it is difficult to automate design. Great engineering talent, for instance, is often wasted making routine decisions on matters such as the size of valves. If some of these routine decisions could be automated, the productivity of a systems engineering team could be increased substantially.

The lack of key sensors can prevent the installation of advanced contol systems. We need better, more reliable, fundamental measurements of the composition and condition of material being processed. We need direct measurements of the quality of final products. For every achievement of an unattended industrial plant, suitable solutions have been available to the problem of directly measuring the quality of the final product.

Another factor that slows acceptance of advanced control systems is the availability of other ways for plants to solve their problems. The company that makes a product is concerned primarily with making enough material that meets the specifications and satisfying its customers; it often has a number of options for solving these problems, only one of which is more automation.

The factors that influence acceptance of advanced automation can be explained by means of a model based on an analogy to the flow of a fluid through two branches in a flow network (Fig. 6). The model mimics the application of resources to solve process problems. These resources may flow along the path toward advanced automation, or they may flow toward other alternatives.

The driving force is the economic perceptions of the industry. The resistance to the application of advanced automation involves such perceived risk factors as the cost and complexity of control, the lack of systems that are reliable and maintainable, the lack of understanding of the process to be controlled, the lack of qualified, experienced technical people, and the conservatism and inertia of the industry. The other alternatives toward which resources may be directed include adopting a different process, using a new raw material, changing the business plan, finding an alternative market for a product, and negotiating a revised labor contract.

Table 2. Major advances in industrial process control.

Era	Measurement	Control	Communication and display
1900 to 1915	Indicating thermometers and pressure gauges	Manual control; off-on controllers; pneumatic valve actuators	Local indicators
1915 to 1930	Orifice plate used for flow measurement	Pneumatic proportional controller; reset mode added	Field-mounted recorders
1930 to 1945	Laboratory analysis of product quality; recording viscometer	Adjustable controller gain; derivative action added	Pneumatic transmission; centralized control rooms beginning to be used
1945 to 1960	On-line analytical measurements; *p*H measurement; gas analyzers	Electronic controllers; first computer control system installed	Miniaturized instruments; graphic panels introduced; electronic transmission used
1960 to present	Specialized sensors; on-line chromatograph	Feedforward control; direct digital control; supervisory control	Cathode-ray tube console for display; computer hierarchies

Conclusions and Future Directions

The major driving force for acceptance of advanced automation will continue to be the perception of economic benefits by the users. As the cost of control hard-

ware decreases, the economic forces increase. The resistance is steadily being reduced by new technology and by education. Therefore, we can expect a continued evolutionary introduction of advanced control in industry but no revolutionary change.

Research and development should focus on reduction of the resistance. Work should be aimed at systems that have increased reliability and maintainability, reduced cost and complexity, require less expertise to implement, are explainable to users with limited sophistication and knowledge, and can be implemented in an evolutionary fashion. There is need for education of potential users and for increased standardization whenever possible.

These are exciting times for workers in the field of industrial process control. There are great opportunities for improved operation of industrial plants by means of advanced automation. The revolutionary developments in electronics now becoming commercially established are apt to greatly accelerate the evolutionary change in process control in the next 5 to 10 years.

References and Notes

1. *Business Week* (No. 2309), 68 (8 December 1973).
2. C. L. Smith and M. T. Brodmann [*Chem. Eng.* **83**, 129 (21 June 1976)] note that in some industry sectors, batch production is finding renewed interest, because it provides the flexibility to change production rates and grades easily. With rising costs of transportation, decentralized multiproduct batch plants may become more attractive than large continuous, single-product plants. If so, more complex tasks will be required from process control.
3. J. C. Maxwell, *Proc. R. Soc. London* **16**, 270 (1868).
4. N. Minorsky, *J. Am. Soc. Nav. Eng.* **34**, 280 (1922).
5. H. Nyquist, *Bell Syst. Tech. J.* **11**, 126 (1932).
6. H. L. Hazen, *J. Franklin Inst.* **218**, 279 (1934).
7. A. Callender, D. R. Hartree, A. Porter, *Phil. Trans. R. Soc. London, Ser. A* **235**, 415 (1936).
8. P. A. Holst [*Simulation* **17** (No. 3), 131 (1971)] discussed the historical origin of the electronic analog computer.
9. S. J. Mason, *Proc. IRE* **41**, 1144 (1953).
10. V. G. Boltyanskii, R. V. Gamkrelidze, L. S. Pontryagin, *Dokl. Akad. Nauk. Uzb. SSR* **110**, 7 (1956).
11. R. E. Kalman and J. E. Bertram, *Trans. Am. Inst. Electr. Eng.* **77**, (part 2), 602 (1958).
12. R. E. Kalman and R. S. Bucy, *J. Basic Eng.* **83D**, 95 (1961).
13. In a pneumatic control system, signals are transmitted by means of air pressure in instrument lines connecting elements of the system. In a PID controller, the manipulated variable (such as valve position) is proportional to a combination of the value, integral, and rate of change of the error (difference between desired and measured value) in the variable being controlled.
14. G. L. Farrar, *Oil Gas J.* (1959 through 1976).
15. E. J. Kompass, *Control Engineering* **24**, 17 (January 1977).
16. W. E. Vannah and L. E. Slater, *Ind. Eng. Chem.* **48**, 1000 (1956).
17. L. S. Yoxall, *Trans. Soc. Instrum. Technol.* **13**, 238 (December 1961).
18. J. McMillan, *Control* **10**, 512 (1966).
19. A. S. Foss, *AIChE J.* **19**, 209 (1973).
20. V. W. Weekman and W. Lee, *ibid.* **22**, 27 (1976).

21. I would like to acknowledge helpful discussions on this subject with a number of people at the Foxboro Company where I serve as a consultant.

Electronics and National Defense: A Case Study

G. P. Dinneen and F. C. Frick

As large-scale integrated circuits become available to take electronics into the homes and the daily lives of a larger public, we may see a diminution in the role that national defense will play in shaping the future growth of this technology. Up to now, however, that role has been crucial; of all the major ingredients of "the continuing revolution in electronics," only the transistor originated without the stimulus of military requirements and the support of military funding (*1*). At the same time, electronic innovation has bolstered our defenses, stimulated new concepts of military tactics and operations, and, when indulged in by others, presented major challenges to our strategic position among the nations of the world.

It is not possible to examine all these complex interactions in a short article, and we have accordingly chosen to concentrate on a particular case—the problem of aircraft surveillance and warning. We have picked this example because it is one that is most familiar to us, and also because it typifies that aspect of defense, command, and control, which is the special province of electronic technology.

Although the scientific principles that underlie modern electronics were reasonably well understood in the 1930's, it was the urgent and uniquely military requirement for early warning of possible air attack that led to the development of radar and with it the large body of associated technology that became generally available in 1947 through publication of the 28-volume Radiation Laboratory Series. The Radiation Laboratory had been established at the Massachusetts Institute of Technology in 1940 to investigate the exploitation of microwave frequencies for radar. This was considered to be a highly speculative venture at the time, but it proved to be remarkably successful in extending the usable frequency spectrum by some three orders of magnitude. Power sources were developed that could deliver several megawatts of peak power at 3000 megahertz, and kilowatts up to 24,000 megahertz. The short pulses required for good radar range resolution necessitated the design of equipment and components with megahertz bandwidths rather than the few kilohertz required for voice radio circuits. Other system needs stimulated developments that ranged from the theory and design of servomechanisms to techniques for precision engineering of large structures. Included were the pulse techniques and the cathode-ray tube and delay line storage devices that supported early electronic computer development, as well as the circuitry and components that made possible the immediate, explosive growth of television after World War II.

Only 3 years after the dissolution of the Radiation Laboratory, the problem that had stimulated the development of radar returned in a new form to force-feed the evolution of the digital computer and the digital technology that would complement the microwave technology developed during World War II. In August 1948, years earlier than expected, the Soviet Union exploded an atomic device, and, for the first time in history, the United States was forced to consider the possibility of a devastating air attack. Not only was there virtually no active air defense in being, but it was clear that early warning and raid tracking at the level of capability developed in World War II was not sufficient to meet a threat

The authors are, respectively, director and assistant to the director, M.I.T. Lincoln Laboratory, Lexington, Massachusetts 02173.

Fig. 1. WHIRLWIND I: 14,500 vacuum tubes; 1024-word (16-bit) internal memory; operating speed, 20,000 instructions per second; computer volume, 4400 cubic feet; power required, 200 kilowatts. Shown here are the electrostatic storage and, from left to right, J. W. Forrester, N. H. Taylor, J. A. O'Brien, C. L. Corderman, and N. L. Daggett.

wherein a single aircraft could inflict more damage than could be expected from mass bombing by fleets of conventionally armed planes (*2*). Furthermore, if the defense was to have sufficient time to intercept high-speed jet aircraft, it would need to employ large, long-range ground radars, and such radars could be bypassed by low-flying aircraft that would remain undetected in many regions.

The solution that was eventually proposed was to use a large number of radars, which together could provide both high- and low-altitude coverage. These were to be netted through "data analysers," which were expected to coordinate all the data from the individual radars and, in addition, it was believed, could control interceptors "at no additional cost in complexity" (*3*).

This admittedly hazy notion of what the central computer would be like resolved itself in prototype as the MIT WHIRLWIND (Fig. 1), which had been under construction with U.S. Navy and U.S. Air Force support since 1947 and was the first computer designed with the heavy emphasis on reliability and speed that would make it suitable for real-time applications. It was clear, however, that the technology that was available in the early 1950's was not satisfactory for the air defense application. In 1951, J. W. Forrester and R. R. Everett, leaders of the WHIRLWIND project, wrote that "no existing digital computer is more than a laboratory model with respect to the present problem. . . . As now foreseen, . . . the proposed machine will be based on transistors as the active elements in the computing circuits and three-dimensional arrays of magnetic cores for high-speed internal storage . . ." (*4*).

As it turned out, the Air Defense Computer (AN/FSQ-7) was not transistorized. Transistors with the uniformity, reliability, and speed that would make them suitable for the highest speed computers did not become available until the announcement of the surface barrier transistor in 1954 (*5*). The magnetic core memory, however, did prove successful and did remove the very serious constraints on computer development that had been imposed by the unreliability and cost of other internal information storage devices, particularly the electrostatic storage tube.

A core memory operated in the Lincoln Memory Test Computer in 1953 and was later transferred to WHIRLWIND. By 1955, core memories were commercially available in the IBM 705 and, shortly thereafter, in the IBM 704, the two machines that for some time dominated the large-scale computer field. In the meantime, the requirements of the Semi-Automatic Ground Environment (SAGE) system led to the demonstration of a number of new roles for the machine that went far beyond any previous concepts of the function of an electronic "numerical calculator," and included display generation, data management, simulation, and real-time, on-line operation.

Development of the central computer was not the only critical problem in design of the air defense system. The deficiencies of early radar (*6*) had forced the British to net together the stations of the "Home Chain," which formed the backbone of their early warning and tracking system for the greater part of the war. This was never a very satisfactory arrangement. Detection, identification, and tracking were all performed manually and coordinated centrally through voice and teletype reports to form a picture of the air situation. Serious delays and errors appeared to be unavoidable, and, with the advent of high-speed aircraft and nuclear weapons, such performance would be intolerable. It was apparent that some means other than "voice telling" would have to be devised to transmit radar data directly to the control center. It was also apparent that microwave relay equipment then available was expensive and not reliable over long distances, and that raw radar data would, in any event, overwhelm the capacity of the central computer.

The solution proposed was to use telephone lines that existed in all populated areas of the country, and to accept the fact that the narrow bandwidth associated with these lines would necessitate extensive filtering and processing of the radar data prior to transmission. Ultimately, this proposal required the installation of massive (up to 1000 vacuum tubes), special-purpose, digital processing equipments at each radar site. In terms of future impact, perhaps the most important outcome was that it led the American Telephone and Telegraph Company to inaugurate a generally available, high-quality, digital data service (*7*).

In the meantime, other military requirements initiated other developments. Missile guidance systems were an early source of support for integrated circuit development; requirements for satellite-tracking radars have supported the development of surface acoustic-wave technology and charge-coupled devices, as well as modern signal processing techniques, including pulse compression and matched filtering. The need for reliable military communications has led to the exploitation of tropospheric and ionospheric scatter, to spread-spectrum and noise modulation systems, and to the development of satellite communication to small, mobile terminals. On-line computer time-sharing, computer net-

works, and computer graphics were all initially supported by the military. Finally, we should mention such significant second-order developments as radio and radar astronomy, microwave spectroscopy, and instrumentation for earth-resources satellites and for modern health care, all of which are heavily dependent on concepts and components derived from military electronics.

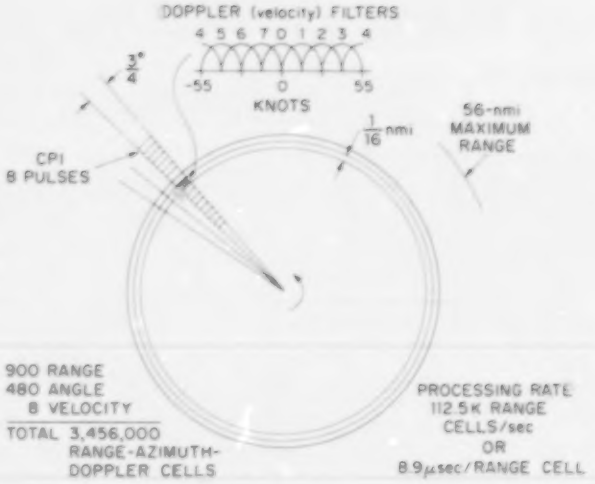


Fig. 2. Signal processing by the Moving Target Detector (MTD) for an airport surveillance radar (ASR-7).

Impact on Defense

So far, we have primarily noted the impact of defense requirements on the growth and evolution of electronics. There is, of course, a reciprocal effect. Prior to World War II (*8*)

> . . . military requirements appeared to be reasonably well understood and straightforward. Ground forces and navies were the product of long experience and gradual evolution. The increasing mobility of the tank and the long-range firepower of the aircraft were beginning to reshape the face of war, but even they were evolutionary platforms and had undergone trials in World War I and subsequent conflicts. Force planning could be, and was largely traditional and incremental although occasional and annoying innovators such as airpower and tank enthusiasts threatened to disturb the customary patterns of warfare by suggesting novel uses for newer military instruments.

Now, however, conditions have changed dramatically. Not only must we support three basic types of forces—strategic nuclear, theater nuclear, and conventional—but, as the technology of weapons, missiles, and jet aircraft which has given us almost limitless firepower and mobility becomes available to others, the classic measures of military strength must be balanced by how quickly and precisely this strength can be controlled. To quote M. R. Currie, director of Defense Research and Engineering: "Absolutely central to our strategic posture is the effectiveness of our command, control, warning, and surveillance systems. . . ." In the tactical area, "the development of integrated intelligence and target engagement systems represents the great challenge and opportunity of the next decade for DoD" (*9*).

Electronics has already, of course, given us initiative in these areas. It has supplied means for instantaneous communication with any part of the world, the capability for continuous planet-wide surveillance and a high degree of precision in weapon guidance. It is difficult, furthermore, to see how we could manage our $125-billion investment in weapon systems or assign and support 2 million uniformed personnel in more than 100 separate locations throughout the world without computer assistance and electronic information storage.

In principle, electronic information processing affords the capability that is required to coordinate our multiforce military structure and to carry out an increasingly complex defense mission. In fact, however, the multiplicity of new sensors, new information sources, and new communication capabilities tends to produce an overwhelming flow of raw data rather than the refined information that recognizes a threat, identifies targets, permits reliable tracking of a mobile opponent or, more generally, will support the degree of integration and automation that is necessary for rapid, effective control of modern military forces (*10*). It still remains to be seen whether the continuing revolution in electronics carries within it means for control and utilization of the massive flows of data which it has, in large part, spawned.

This surfeit of data appears in many areas and will not be susceptible to any single remedy. We believe, however, that there is emerging a qualitatively new capability for local information processing, derived from large-scale integrated circuitry, that carries intimations of the future resolution of many of these problems. In order to illustrate this thesis, we propose to describe in some detail recent results that have been obtained in connection with the persisting problem of aircraft surveillance and tracking. We have picked this particular development rather than comparable achievements in speech processing or radar, infrared, or visual imaging, because it is the clearest example that we know of the application of this newly available capability to the extraction of meaningful information from a vast and cluttered flow of data. We also find it interesting that the military problem of air surveillance, which has initiated so many important developments in modern electronics, is one of the first problems that may ultimately be solved through advances in that technology.

It was recognized early (*11*) that interceptions of high-speed aircraft could never be fully successful without full automation that included automatic detection and tracking, but the air defense systems of the 1950's never achieved this goal. Studies (*12*) have shown that for an automatic tracker to perform adequately in a noisy environment, the probability of target detection should be .8 or higher per scan, and the false alarm rate should be less than 10^{-6}. Despite improvements in radar resolution, moving target indication (MTI) and clutter mapping, the basic data input remained vulnerable to countermeasures, contaminated by ground returns, weather, birds, and rf interference, and the requisite performance standards could not be regularly achieved. Furthermore, to track maneuvering aircraft successfully, it is important that there be a low scan-to-scan correlation of detection dropouts. Because of these difficulties, even the most sophisticated computerized tracking algorithms have needed skillful human assistance to recover from sporadic loss of data and to track through the clutter that was inevitably present.

To achieve such low false alarm and high detection rates, it is necessary to examine the area scanned by the radar by dividing it into a large number of range-azimuth–Doppler cells (see Fig. 2). A backscatter return in such a cell is reported as a target only if the magnitude of the return is above the momentary

threshold of the cell. If this threshold is set too high, the detection rate will suffer; if it is set too low, there will be a large number of false alarms triggered by various forms of noise and clutter. An optimal system will require variable (adaptive) cell thresholds that are conditioned by local context—the returns observed in the cell on prior scans and returns observed in spatially adjacent cells.

This represents a formidable data processing task which, in the particular application considered here, is accomplished through digital processing of the converted outputs of the video detector. As shown in Fig. 2, the region under surveillance is divided into 480 azimuth intervals, each 3/4° wide (one-half beam width), extending approximately 56 nautical miles in range. This range extent is divided into 900 range cells by sampling the return at a 1.3-megahertz rate, equivalent to 1/16 mile in range. The area is thus divided into 432,000 range-azimuth cells.

In each 3/4° azimuth interval, called a coherent processing interval (CPI), eight pulses are transmitted at a constant pulse repetition frequency (PRF) of approximately 1 kilohertz. On reception, eight complex digital samples are collected in each range-azimuth cell, and these are processed to produce eight Doppler (velocity) filters. This divides the area into 3,456,000 range-azimuth–Doppler cells, each of which is given its own threshold.

For targets with zero radial velocity, the detection threshold in each range-azimuth cell is a weighted average of the scan-to-scan history, which is stored in a digital map of the ground clutter. The map value is built up in a recursive manner by adding 1/8 of the output of the zero velocity filter on each scan to 7/8 of the value stored in the map. Thus, as rain moves into the area or as propagation conditions change, the map values also change. The value stored in the map is multiplied by an appropriate constant to set the threshold for the zero velocity filter.

Unlike ground clutter which has a constant spectral width centered at zero velocity, precipitation returns have a spectral shape which is set by the wind field and varies in width as well as average velocity. If the PRF of the radar is changed on successive CPI's, the very narrow Doppler spectrum of the aircraft will appear in different Doppler filters with each PRF. Except when the target's radial velocity is exactly that of the rain, aircraft returns will appear in some Doppler gate without any rain return on at least one of the two PRF's as the antenna beam scans past. It is thus possible to detect aircraft whose cross section is many decibels below the radar cross sec-

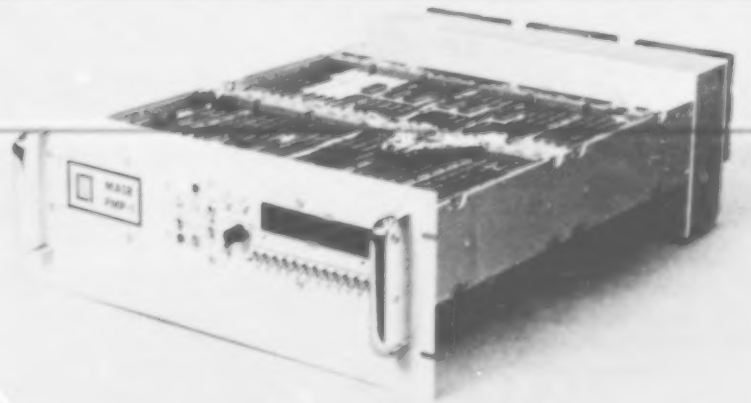


Fig. 3. Parallel Microprogrammed Processor (PMP): 3000 integrated circuits including 1024-word (24-bit) and 2048-word (60-bit) internal memories; operating speed, 25×10^6 instructions per second; computer volume, 1.8 cubic feet; power required, 160 watts.

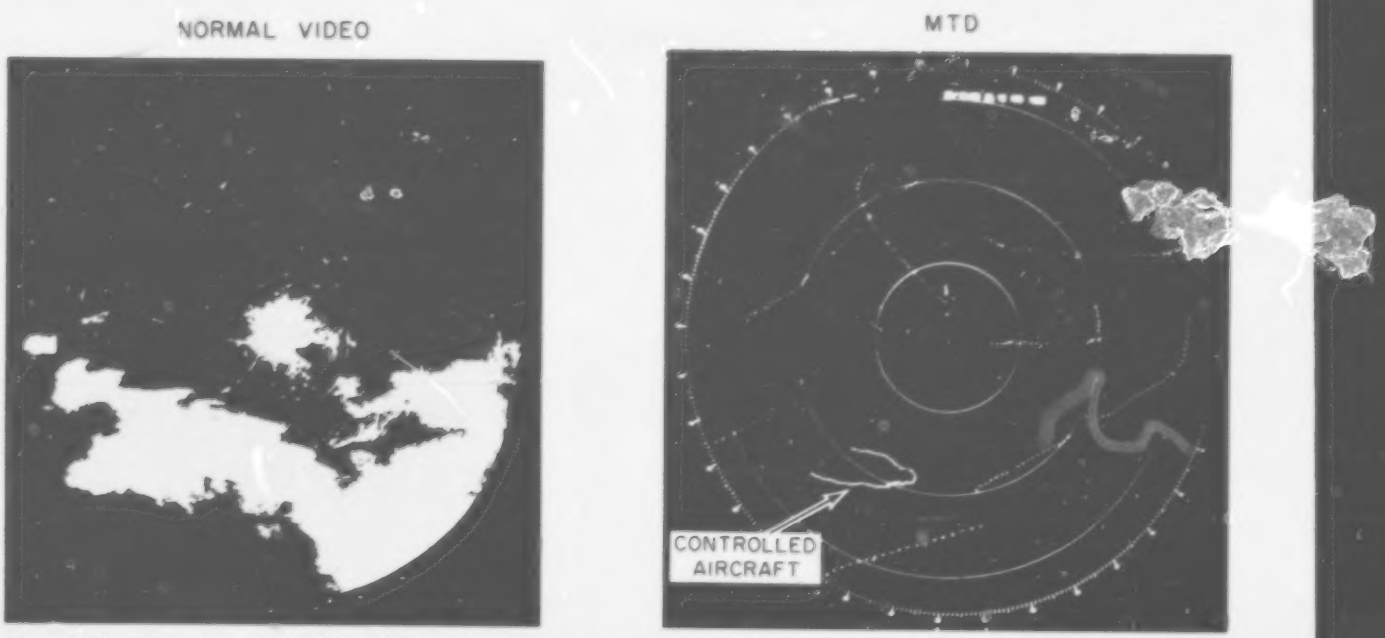


Fig. 4. Performance of the Moving Target Detector (MTD) in heavy precipitation and ground clutter. Normal video and the tracked aircraft outputs are shown. Notice the absence of false returns and the continuous tracking even of aircraft with zero radial velocity. The controlled aircraft is a single-engine Piper Cherokee.

tion of the weather return. For targets with a radial velocity, the detection threshold is set as a weighted average of returns in 16 range cells, eight on either side of the cell of interest. Blind speeds introduced by traditional MTI cancellers are also eliminated by changing the PRF every coherent processing interval.

Prototype equipment that will carry out these operations is pictured in Fig. 3. It is called the Parallel Microprogrammed Processor (PMP) and consists of a control unit and one or more identical processing modules (PM's), each containing a processing element and a data memory. Each processing element performs all the calculations on the data in its own data memory while the control unit, under microprogram control, sequences through the program in its control memory, providing instructions and data memory addresses to each PM. The operations of data collection, velocity filtering, and adaptive thresholding are performed for each range cell. The signal processor can, therefore, be a pipeline device that processes the range gates sequentially, or it can be, like the device shown in Fig. 3, an array of slower processors, all running in parallel and each processing only the data in a portion of the range window.

The results of tests performed with the Moving Target Detector (MTD), a hardwired predecessor of the equipment pictured here, have been dramatic. An example is presented in Fig. 4, and equally good performance has been observed in the presence of radio-frequency interference.

Of particular interest when considering the broader implications of such a system is the observation that the false alarm rate is so low that the entire radar output can be reliably transmitted over a telephone line, and handover or correlation of tracks among a number of radars becomes primarily an exercise in coordinate conversion. This not only relieves the communication load, but it unburdens the central control system to operate at a different level of aggregation; for instance, to integrate radar tracking with the output of other types of sensor as might be required for full battlefield surveillance.

The technology involved—highly integrated, high-performance digital circuitry—has made it possible to carry out a massive, real-time data processing task that would not have been contemplated even a few years ago (*13*); and it has made such processing economical enough to associate this capability with each sensor, thus taking advantage of local contextual factors that, as in the example shown, are often the only means by which it is possible to extract meaningful information from the clutter in which it is generally embedded.

The low cost of digital data processing equipment also encourages duplication of circuitry for improved reliability and eventually, perhaps, completely unattended operation. The performance of the present system, for example, degrades gracefully because each PM processes only a portion of the full range window, and the failure of a PM means that the full range window continues to be processed except for that particular portion covered by the failed PM. In addition, there is a potential for self-repair in the sense that, if a PM is failing and this failure is detected by a diagnostic program running concurrently with the real-time processing program, then the processing load carried by the failed PM could be automatically switched to a standby PM. Coupled with evolutionary advances in solid-state components, nonmechanical switched-array antennas, automatic monitoring, and data-remoting equipment, such a system could operate for months at a time with no on-site maintenance or other manning.

Political and military imperatives will continue to push the technology for faster processors, larger memories, more communications, heuristic processing, worldwide conferencing, rapid and automatic translation of languages, and other electronic capabilities to ensure reliable, informed, and secure control of our military forces. Concurrently, the innovative use of the technology can have a significant effect as a "force multiplier" and in achieving economy of manpower on the battlefield.

It is expected that the shared interest of industry and the public generally will make them partners in the development of these capabilities; but if history is a guide, considerations of national defense will continue to provide a major impetus for the invention and innovation that is necessary to maintain a continuing revolution in electronics.

References and Notes

1. Even here the military presence has been significant in the continuing development of solid state electronics. It was major U.S. Air Force funding that led to the development of the silicon transistor in 1954. The need for simplified field maintenance and the stringent power and weight constraints on airborne equipment motivated Air Force support of "molecular electronics," an unfortunately named project that was nevertheless one of the key initiating events in the development of integrated circuits [P. E. Haggerty, *Proc. IEEE* **52**, 1400 (1964); M. F. Wolff, *IEEE Spectrum* **13**, 45 (1976)].
2. J. R. Killian, Jr., and A. G. Hill, *Atlantic Monthly*, November 1953.
3. Memorandum on Activities of the Air Defense System Engineering Committee, 7 April 1950, cited in *Lincoln Laboratory Quarterly Progress Report*, 1 June 1952, p. xii. The same memorandum concludes: "The type of mechanism best suited to perform this function is called a digital computer. It is similar in general idea to an IBM punched card bookkeeping machine but thousands of times faster . . . it is expected that about 100 data analysers will be required to cover 2,000,000 square miles at a price of about $500,000 per unit."
4. Memo to Project Lincoln Steering Committee, 5 October 1951.
5. J. L. Mitchell and K. H. Olsen, *Proceedings of the Eastern Joint Computer Conference* (December 1956), pp. 93–101.
6. L. N. Ridenour, Ed., *Radar System Engineering* (McGraw-Hill, New York, 1947), pp. 175 and 228.
7. The telephone company was a reluctant partner. In the early 1950's, it was not easy to see much general demand for the transmission of digital information.
8. D. H. Rumsfeld, *Report of Secretary of Defense to the Congress*, 1976.
9. M. H. Currie, *Department of Defense Program of Research, Development, Test and Evaluation, FY 1977* (February 1976).
10. "Military radioelectronics assures not only the use of missiles, anti-missiles, and other technical means of combat, but also reconnaissance, the control of troops, forces, and weapons as a whole. It is the basis of the solution of the problem of complex automation of the processes of staff activity. Without complex automation, effective command of the armed forces and consequently their successful use in a modern war will be impossible," V. D. Sokolovskiy, in *Soviet Military Strategy (1968)*, H. F. Scott, Ed. (Crane Russak, New York, 1975), p. 198.
11. Letter, Director of R & D, Deputy Chief of Staff to Commanding General of the Air Defense Command, 16 January 1948, quoted in *History of the Electronics Systems Division*, AF Support Command Historical Publication 65-30-1 (Government Printing Office, Washington, D.C., 1964).
12. The work on radar data processing reported here was carried out under the direction of C. E. Muehe, MIT Lincoln Laboratory, and we have borrowed heavily from reports given by Muehe and members of his group. It is interesting that the initial development of the MTD radar processor was supported by the FAA for "air traffic control" rather than the military for "aircraft control and warning."
13. Some substance is given to this comment by a comparison of the PMP with Illiac IV. Illiac IV was designed in the mid-1960's to be the most powerful computer that the state-of-the-art would permit. It is the largest single machine in operation, and it is capable of executing approximately 200 million instructions per second [see W. J. Bouknight, S. A. Denenberg, D. E. McIntyre, J. M. Randall, A. H. Sameh, D. L. Slotnick, *Proc. IEEE* **60**, 369 (1972), and Howard Falk, *IEEE Spectrum* **13**, 65 (1976)]. An expanded PMP, with five processing modules, would be capable of executing approximately 125 million instructions per second.
14. This article was prepared at Lincoln Laboratory, a center for research operated by MIT with the support of the Department of the Air Force under contract F19628-76-C-0002.

Some Recent and Future Automotive Electronic Developments

Trevor O. Jones

There have been many conferences and much has been published (*1*, *2*) on the subject of automotive electronics, and now, with the introduction of a microprocessor to control the ignition timing of the 1977 Oldsmobile Toronado, a new age of automotive electronics has been unveiled. This first application of a microprocessor comes 25 years after the introduction of the first electronic control system on the automobile. This was the headlight control, which was introduced in 1952 and used the analog vacuum tube technology of the day. Over the past 25 years, 17 automotive electronic systems have been introduced, excluding communications and entertainment equipment. Sixteen of them remain and are shown in Table 1 in order of their introduction. This list excludes seat belt interlocks and traction control, which are no longer offered.

With the rapid advances being achieved in solid-state technology, increased packaging density, and lower cost and improved reliability of the systems, it is understandable why there is considerable interest in the application of this technology to the automobile. The stimulation provided by new technologies and the pressures of excessive and, at times, unnecessary governmental regulations have accelerated the automotive application of solid-state technology.

One measure of this increase in activity is the projected electronic content as a percentage of automobile cost. One source (*3*) predicts that by 1985, 10 percent of the cost of an automobile may be represented by electronics. That would represent a large electronic sales potential in excess of $10 billion annually, not including entertainment and communications equipment.

The size and complexity of an automobile electrical system makes it a prime candidate for the application of contemporary digital systems control technology. In addition to the 16 electronic subsystems previously mentioned (Table 1), an automobile can have 420 meters of cable and 83 switches which may be replaced in the future by solid-state devices. There can be 14 electric motors and 69 lamps. Also, there are 27 sensors and 27 fuses and breakers to protect the circuits, which are tied together with up to 108 connectors.

Emissions and Fuel Economy

Two government regulations regarding emissions and fuel economy will have a significant impact on the continuing application of electronics to the automobile. Neither of these regulatory requirements can be satisfied with current technology. The 1978 statutory emission control standards (*4*) promulgated by the Environmental Protection Agency (EPA) require automobiles to be designed so that they do not exceed the exhaust emission values shown in Table 2 during the first 50,000 miles of vehicle operation (*5*). As emission standards become more stringent, a reduction in fuel economy may be anticipated. More stringent emission standards are in direct conflict with the intent of Public Law 94-163 (*6*), which requires each manufacturer to improve fuel economy and to have a production-weighted average fuel economy of 18 mile/gallon in 1978, 20 mile/gallon by 1980, and 27.5 mile/gallon by 1985.

Fuel economy improvements are being achieved first by making vehicles smaller and lighter, and second by improving engine and power train efficiency and performance. It is in the second area that electronics is expected to be applied; it is also in this area that strict attention must be paid to meeting the stringent emission standards.

Four electronic systems associated with engine control have been introduced or announced in the past 2 years. This is an addition to solid-state high-energy ignition systems which are essentially standard on American automobiles. Of the four, the most advanced to be introduced is the Delco-Remy MISAR system (microprocessed sensing and automatic regulation), which precisely controls spark timing for all conditions of load and speed consistent with drivability and emission control requirements (*7*). This system is being offered on the 1977 Oldsmobile Toronado and is particularly significant because it is believed to be the first application of a microprocessor to the automobile. A simplified diagram of the MISAR system is shown in Fig. 1. The input signals are crankshaft position, manifold vacuum, coolant temperature, and reference timing. The microprocessor is equipped with metal oxide semiconductors (MOS) and large-scale integration (LSI) technology and its read-only memory (ROM) has a capacity of 10,240 bits. The MISAR system can be readily expanded to other engine control functions because of unused ROM capacity. The system permits the engine to be tuned to improve fuel economy and drivability. It has improved the fuel economy on the Oldsmobile Toronado by approximately 10 percent while maintaining excellent drivability. Chrysler introduced a lean-burn spark timing control system utilizing analog technology on their large- and intermediate-size vehicles in 1976.

Electronic fuel injection (EFI) systems can provide improved starting and drivability in relation to carburetor systems. One such system is standard equipment on the Cadillac Seville. This was introduced in 1975 and is optional on other Cadillac models. The current EFI system is basically analog, but is a near-term candidate for the application of microprocessor technology.

Ford Motor Company recently announced (*8*) an electronically controlled dual displacement engine (DDE). The engine will be introduced in the next 2 years on light trucks, and it is expected to improve fuel economy by at least 10 percent and perhaps up to 20 percent. This improvement is achieved by deactivating half the cylinders when the vehicle speed exceeds 45 mile/hour or when it is decelerating to 25 mile/hour. Upon acceleration, all cylinders are activated. The electronic engine control system computes commands based upon engine temperature, transmission gear, engine vacuum and speed, and throttle opening.

The author is director of General Motors Proving Grounds with headquarters at Milford, Michigan 48042.

A closed-loop knock-limiting system was developed by General Motors in 1976 which may allow improvements in fuel economy while permitting an engine to operate at acceptable knock levels on current 91 octane fuels. The system was designed to control ignition timing only under knocking conditions so that optimum timing may be used under all other conditions to improve fuel economy while satisfying exhaust emission constraints.

A sensor determines engine knock and provides information to an electronic logic system which retards timing to acceptable knock levels. An acceptable system must control knock under all operating conditions of speed, load, temperature, humidity, or altitude without significant degradation in economy, drivability, performance, or emission levels. A low-cost device to sense knock at trace levels in all cylinders without mechanical or electrical interference is required.

Catalysis is the only known method of achieving the 1978 statutory emission standards; however, the technology has not matured sufficiently to meet the statutory standard for 50,000 miles, and a change in the law is essential if the industry is to continue to produce acceptable automobiles. Systems which simultaneously control hydrocarbon (HC), carbon monoxide (CO), and nitrogen oxide (NOX) emissions with a single catalyst require the air and fuel mixture to be precisely controlled to the stoichiometric ratio (14.6 : 1), at which point good catalyst efficiency is obtained (Fig. 2) for the three regulated by-products of combustion. This level of control is achieved with a system in which the oxygen level in the engine exhaust is used as the sensing parameter. The oxygen transducer is essentially a battery comprised of a solid electrolyte of zirconium dioxide with platinum-plated electrodes on either side. At the stoichiometric ratio and at exhaust gas temperatures, the zirconia sensor provides excellent signal characteristics (Fig. 2) which coincide with the optimum catalyst efficiency for the three constituents of interest. A simplified block diagram of the closed-loop system is shown in Fig. 3. Closed-loop emission control systems may be used on both EFI and carbureted engines if a three-way catalyst is used, and considerable effort is being expended by the automobile industry to develop such systems.

Alternative power plants, such as the Stirling cycle engine, gas turbine engine, and electric engine, are post-1980 developments (9).

Table 1. Contemporary automotive electronic subsystems.

Headlight control	Wheel-lock control
Alternator rectifier	Clock
Voltage regulator	Intrusion alarm
Tachometer	Air-cushion control
Cruise control	Electronic fuel injection
Electronic ignition	Lamp timing control
Climate control	Spark timing control
Windshield wiper control	Electronic digital displays

Table 2. The exhaust emission values that must not be exceeded during the first 50,000 miles of vehicle operation, according to the 1978 statutory emission control standards (4).

	Emissions (g/mile)		
	HC	CO	NOX
1978 Standards			
Federal	0.41	3.4	0.4
1977 Standards			
Federal	1.5	15	2.0
California	0.41	9	1.5

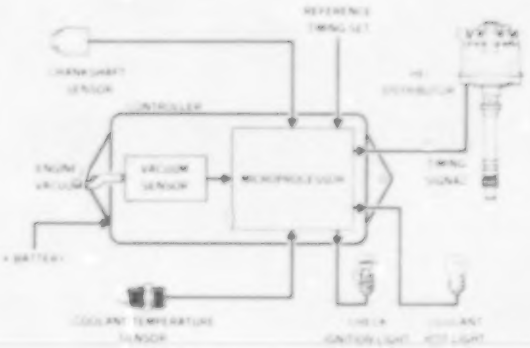


Fig. 1. The Delco-Remy MISAR system which controls spark timing for all conditions of load and speed.

Truck Systems

There have been three recent applications or announcements of electronic systems for trucks. Government standards (10) for buses, trucks, and trailers with air brakes were sufficiently stringent that manufacturers resorted to the use of wheel-lock control systems in order to meet the performance levels specified. The basic problem was that initial requirements were too stringent, and insufficient development time was allowed for by the government regulators in spite of substantial and continuing pleas on the part of the vehicle manufacturers. The resulting unacceptable reliability of bus wheel-lock control systems prompted the government to suspend stopping-distance requirements for these vehicles. The wheel-lock control systems offered today are all analog in nature, and a microprocessor will no doubt be used for the control unit in the future.

Ferranti announced (11) the development of a flexible load-weighing system for trucks. The system derives weight data from a strain gauge attached to each axle. The load on each axle is computed and displayed as is the total vehicle load. The system can detect inadvertent overloads which can result in higher maintenance cost and possible fines, and can indicate underload situations which, if corrected, can improve the economics of hauling.

General Motors has developed a low tire pressure warning system for trucks, tractors, trailers, and buses. The system consists of four major components: a temperature-compensated pressure switch, an actuator-sender that replaces one wheel mounting bolt, a stationary

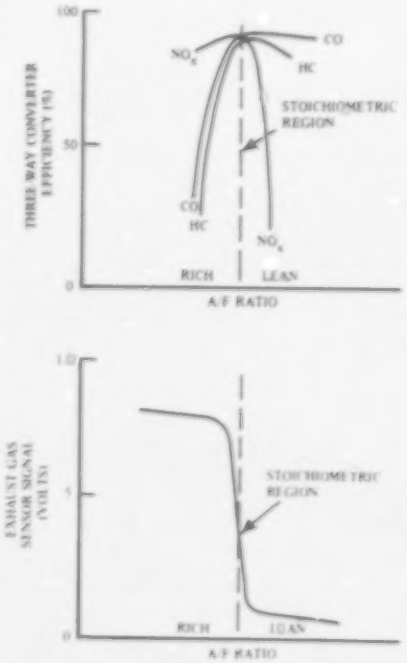


Fig. 2. Typical characteristics of a three-way converter and an exhaust gas sensor.

sensor behind the mounting bolt, and a display panel. Low tire pressure lights a panel indicator and a "pop-up" button in the wheel actuator provides out-of-cab detection. The warning continues until the tire has been inflated and the "pop-up" button reset. The low tire pressure warning system maximizes tire life and fuel economy in addition to improving safety and reducing maintenance inspection time.

Displays

With the start of the 1977 model year, General Motors introduced electronic displays integral with an automobile AM/FM radio. The four-digit, seven-segment light-emitting diode (LED) displays four functions on command, namely, date, time of day, elapsed time, and radio station frequency. In addition to LED's, electronic display technology under development around the world includes gas discharge, liquid crystal, electrochromic, and electroluminescence. Cost and demonstrated life continues to present major barriers to expanded use of electronic displays. However, it is anticipated that future automobile instrument clusters will be generated from a single mask by means of these technologies. Prototype d-c electroluminescence instrument clusters have been manufactured from a single mask and demonstrated in vehicles (2). This system displayed both digital and analog data.

Multiplex Systems

As electrical and electronic systems are added to the automobile, it becomes progressively more difficult to meet reliability objectives because of multiple connectors. Multiplex systems have the potential to improve reliability and reduce weight, and they can be designed to be immune to noise. It appears that adequate technology would be available for the use of such systems if their cost were more competitive with current wiring systems. An additional advantage of such a system is the relative ease of introducing diagnostic routines, provided low-cost transducers are available.

An interesting development is the application of fiber optics to automotive multiplexing systems. Advances in fiber-optic control systems during the past 2 years include more efficient infrared light sources, low noise photo diodes, and rugged, efficient fiber-optic cables, many of which are now available off the shelf.

Lightweight, interference-free fiber-optic cables are replacing heavy, complex copper wire bundles aboard military aircraft and ships. Underground communication cables are being replaced with light pipes capable of transmitting conversations over many kilometers. Current fiber-optic technology for data lines can provide a weight advantage of up to 100 : 1 and a cost advantage of 30 : 1 compared to copper lines.

The extensive advantages of this technology include the immunity of the fibers to electromagnetic interference (EMI) and to oil, grease, dirt, and acids, and high-speed information transmission. These advantages are of considerable importance to the automobile engineer. The advent of lower cost components, coupled with the inherent performance advantages of this technology, make it an ideal means of reducing vehicle harness weight and providing EMI-immune systems.

A multiplexed control system with a single fiber-optic cable (see Fig. 4), has been developed to replace the multiwire harness, connectors, and printed circuitry required to interconnect steering column switching functions with vehicle power actuators and lights.

An encoder module, located at the top of the steering column, interprets the following switching commands and digitally multiplexes each command signal: windshield wiper, turn signal, hazard flashers, key reminder, horn, headlamp controls, cruise control. These signals are transmitted as light pulses through the fiber-optic cable to a decoder module located at the base of the steering column. This module decodes individual switch commands and sends a signal to the appropriate power actuator, for example, the windshield wiper motor, horn relay, or lights.

An inexpensive LED on the encoder chip is electrically pulsed. The 32-bit digital words are converted by the LED to light pulses which are transmitted down the cable. A photo diode on the decoder chip receives the light pulses as

Table 3. A list of proposed electronic systems for the automobile. Prototypes of many of these systems have already been made.

Automatic cruise control	Vehicle locator
Automatic brakes	Emergency location transmitter
Radar crash sensors	Multiplex wiring systems
Crash recorders	Low tire pressure indicators
Electrocardiograms	Vehicle load weighing systems
Sleep detectors	Ultrasonic fuel injectors
Alcohol ignition interlocks	Engine knock limiting control
Automatic vehicle guidance	Dual displacement engine control
Road coefficient detector	Closed-loop engine control
Controlled exhaust gas regulation	Electronic displays
Automatic diagnostic systems	Current drain protection
Vehicle blind spot detectors	

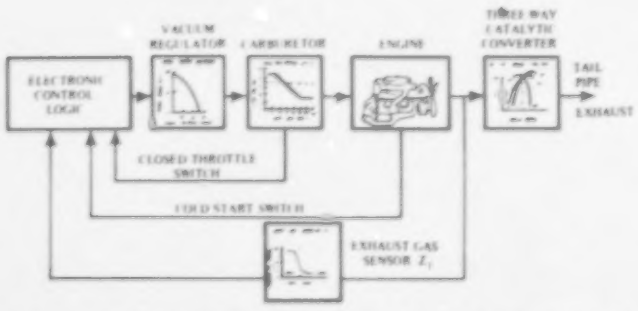
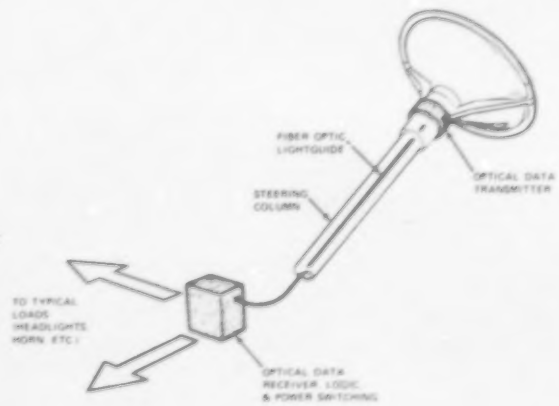


Fig. 3 (left). Diagram of a closed-loop system for use in controlling the air and fuel mixture. Fig. 4 (right). A multiplexed control system with a single fiber-optic cable.


FIBER OPTIC LIGHTGUIDE
OPTICAL DATA TRANSMITTER
STEERING COLUMN
TO TYPICAL LOADS (HEADLIGHTS, HORN, ETC.)
OPTICAL DATA RECEIVER, LOGIC & POWER SWITCHING

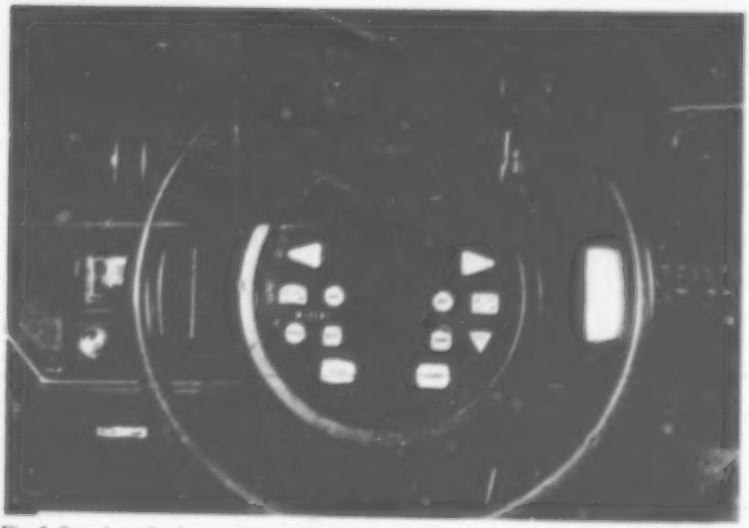


Fig. 5. Steering wheel control panel. The main switches are on a nonrotating platform in the center of the wheel.

Fig. 6. X-band radar signatures.

they emerge from the cable and reconverts them to electrical signals which are recognized by the decoder on individual switch commands.

A 1975 Buick Electra was equipped with a fiber-optic multiplexed steering column system (Fig. 5). Design flexibility was demonstrated by placing major vehicle switching functions on a nonrotating platform in the center of the steering wheel. Small switches located on the platform, or anywhere on the instrument panel and door panels, can replace larger, high-current carrying switches required with conventional systems. This feature provides considerable design flexibility. For instance, a citizen's band (CB) microphone could be mounted in the center of the platform to carry voice signals down the fiber-optic cable to the CB transmitter unit.

Systems of this type are not without problems, however. Several major engineering hurdles must be overcome before optical multiplex systems are applied to an automobile. A low-cost solid-state switch capable of converting low-current switching commands from the decoder module to high-current power drive signals is required. These developments further emphasize the need for automotive systems to be cost-effective.

Long-Range Systems

A list of proposed electronic systems for the automobile would be extensive and exceed 100 items in addition to the current systems shown in Table 1. Some of the advance systems which have been proposed are shown in Table 3. Many of these systems have the potential to become standard or optional automotive equipment in the next decade; several of them are under active consideration and may be seen on automobiles within this decade. In general, the near-term systems suffer from poor cost-effectiveness, generally because of the high cost of the transducers and actuators. While the utility of many of the longer-range systems can be justifiably challenged, the universal need for effective transducers and low-cost actuators cannot. Two of these longer-range systems are of particular interest: alcohol ignition interlocks and automatic radar brakes. If these systems could be successfully developed, they could have a significant impact on improving highway safety. Their implementation would, however, depend on an acceptable level of cost-effectiveness being achieved.

Government studies (*12*) continue to show that the driver is the major contributor to the cause of over 80 percent of vehicle highway accidents. Alcohol continues to be a major cause of highway accidents. Over 50 percent of highway fatalities are due in part to alcohol consumption, and this percentage is expected to rise because of the projected increase in per capita alcohol consumption (*13*). Drivers continue to run into the rear of other vehicles. In another government study (*14*) of rear-end collisions, 45 percent of the drivers in the striking vehicle claimed they never saw the vehicle struck until after the accident. Because of these conditions, it is expected that research will continue in areas aimed at reducing the driver-caused accident, with specific emphasis on research directed toward the ultimate goal of inhibiting the operation of an automobile by an intoxicated or otherwise impaired driver, and toward automatic brakes.

Experiments have been conducted with electronic devices which will prevent a drunk individual from starting his car. The first device developed by General Motors was the breathalyzer (*15*), which rejected approximately 50 percent of potential drivers with a blood alcohol concentration of 0.1 percent. This led to the development of a critical task tracker (*16*), which gave a rejection rate of 75 percent. We are still seeking a device which would be 100 percent effective.

Accidents caused by driver inattention to vehicles stopping or slowing ahead indicate that radar-augmented braking systems may have potential for preventing highway accidents (*17*). The major barrier to designing an acceptable system is the inability of the system to conduct a hazard analysis of potential obstacles. Figure 6 shows X-band radar signatures of three obstacles which are frequently involved in highway collisions. As can be seen, the small stop sign has a signal amplitude considerably higher than that for the 10-inch-diameter tree (*5*). If the system detection threshold were set to include all three obstacles, a large number of false alarms might result which would be unacceptable to the driver.

With the reduced cost of memories, such as charge-coupled devices and bubble memories, the potential exists to store sufficient radar signatures to support a high-speed hazard analysis. The initial application of radar to the automobile is expected to be as part of a cruise control or station-keeping system. If this is successful, it could lead to fully automatic braking systems.

Radar crash sensors for air-cushion restraint systems have been discussed (*18*) in the past. In this application, the

severest hazard analysis situation exists because of the unacceptability of inadvertent or false inflations of the air cushion.

Summary

The automotive industry continues to examine new electronic technologies for their applicability to the automobile. Today, 16 electronic systems can be found on the automobile, and future engine and emission control systems will soon be added. Truck systems of interest include wheel-lock control, vehicle weighing systems, and tire pressure warning devices. Digital electronic displays and multiplex wiring systems are expected to be near-term developments and, on a longer-range basis, it is expected that automatic radar brakes and intoxicated-driver interlocks will receive considerable attention.

References and Notes

1. See, for example, Society of Automotive Engineers, *Automot. Electron.* SAE Publ. No. SP-388 (January 1974); *Convergence 74. International Colloquium on Automotive Electronic Technology* SAE Publ. No. P-57 (Society of Automotive Engineers, Warrendale, Pa., 1974), p. 57; *Convergence 76. International Conference on Automotive Electronics and Electric Vehicles*, SAE Publ. No. P-68 (September 1976); *Automot. Electron.* SAE Publ. No. SP-417 (February 1977); *Automot. Electron. II* SAE Publ. No. SP-393 (February 1975).
2. Institution of Electrical Engineers, *Automob. Electron.* IEE Conference Publ. No. 141 (July 1976).
3. G. F. Villa, in *Proceedings of Convergence 74, International Colloquium on Automotive Electronic Technology* SAE Publ. No. P-57 (Society of Automotive Engineers, Warrendale, Pa., 1974), p. 149.
4. Clean Air Act of 1970, Title 42, U.S. Code Emission Standards for Moving Sources, Chapter 2, Section 202 (Government Printing Office, Publ. 0-445-880, Washington, D.C., 1971).
5. The metric equivalents for units used in the text are as follows: 1 mile = 1.6 km; 1 gallon = 3.7 liters; 1 in. = 2.5 cm.
6. Energy and Policy Conservation Act, PL 94-163 Title 5, "Improving Automotive Efficiency" (Government Printing Office, Washington, D.C., 1976).
7. D. J. Sumanaitis, *Automot. Eng.* 85, 24 (1977).
8. Vehicular Technology Group, *IEEE Vehicular Technol. Group Newsl.*, p. 6 (November 1976).
9. J. J. Brogan, in *Proceedings of Convergence 76, International Conference on Automotive Electronics and Electric Vehicles* SAE Publ. No. P-68 (IEEE, New York, 1976), p. 24.
10. Federal Motor Vehicle Safety Standard 121, "Air Brake Systems" (National Highway Traffic Safety Administration, Department of Transportation, Washington, D.C.).
11. H. F. Dickinson, *Automob. Electron.* IEE Conference Publ. No. 141, p. 33 (July 1976).
12. Institute for Research in Public Safety, "Tri-Level Study of the Cause of Traffic Accidents," Contractor's Report No. DOT-HS-034-3-535-TAC (National Technical Information Service, Springfield, Va., October 1974).
13. Department of Transportation, "The National Highway Safety Forecast and Assessment (A 1985 Traffic Safety Setting)" (National Highway Traffic Safety Administration, Department of Transportation, Washington, D.C., 1975).
14. "Problems of Linear Closure at Urban Intersections," Contract No. CPR-11-4201, Second Quarterly Progress Report (National Highway Traffic Safety Administration, Department of Transportation, Washington, D.C., 1966).
15. T. O. Jones and J. A. Tennant, "Alcohol Impairment Detection by the Phystester—Evaluation Summary," SAE Paper 730093 (Society of Automotive Engineers, New York, January 1973).
16. J. A. Tennant, in *Proceedings of Automotive Safety Engineering Seminar* (General Motors Corporation, Warren, Mich., June 1973), p. 95.
17. D. M. Grimes and T. O. Jones, *Proc. IEEE* 62, 804 (1974).
18. T. O. Jones, D. M. Grimes, R. A. Dork, "A Critical Review of Radar as a Predictive Crash Sensor," SAE Paper 720424 (Society of Automotive Engineers, New York, May 1972).

Electronic Mail

Electronic communication of information is more rapid than conventional transportation of documents.

Robert J. Potter

The concept of an organized mail system is believed to have emerged in about 400 B.C. in the Persian Empire. Queen Elizabeth I issued a proclamation in 1591, which prohibited carrying of mail to and from England except by messengers authorized by the Master of the Posts. In America, the first postal system was authorized by the colonial legislature of Massachusetts in 1639.

Efficiency was gained by faster transmission and, subsequently, by faster distribution at the terminals of the system. The pony express increased the extent of the postal system. Airmail reduced the time of transit. And now our society is in the midst of a revolution, the electronic revolution, which makes the delay in transmission of information negligible.

The term "mail" has come to mean the ordinary physical transport of a document, from one location to another in a sealed and addressed envelope. The electronic revolution has made it possible to replace the transportation of the document with the electronic communication of its information—"electronic mail." The decision to use an electronic mail system is primarily sociopolitical and economic. Today's technology can provide a cost-effective electronic mail system under many conditions.

An interesting analogy can be drawn between the evolution of the telephone (from manual switchboards, to electromechanical switching, and now to electronic switching) and mail systems, which also started with manual sorting and human transportation and delivery; they have progressed to automated electromechanical sorting, rapid air transportation, optical character recognition, and have now begun to use electronics for many of their operations.

Several systems on the market today use electronics for communication of information, that is, forms of electronic mail. The telegraph and the telegram (though fading rapidly because of the newer and more advanced forms of electronic communication) were early means of transmitting information electrically. Teleprinter systems, telex/TWX, and Mailgram, more modern forms of electronic mail, are now used throughout the world. Mailroom and point-to-point facsimile are becoming widely accepted electronic systems for inter- and intracompany mail service. Large interconnected computer centers are used to transmit messages to and from computer terminals. The use of communicating typewriters, word processors, various keyboard terminals, and other electronic devices is predicted to grow rapidly because of their high efficiency in transmitting alphanumeric characters.

Kinds of Mail

The mail can be divided into five major types: (i) transactions, (ii) advertising, (iii) correspondence, (iv) magazines and newspapers, and (v) merchandise in order of decreasing volume (*1*). Much mail contains financial transactions such as

The author is the president of the Office Systems Division of Xerox Corporation, Dallas, Texas 75247.

statements, invoices, check payments, and other documents (Fig. 1).

An analysis of the mail to and from the three main sources and recipients, businesses, government offices, and households, is shown in Table 1. Some weights and costs of ordinary mail are shown in Table 2.

The bulk of business and government communications can be split between alphanumeric and graphic information. Words and numbers constitute 90 percent; 10 percent is graphic (*2*). The graphic portion consists primarily of special forms, such as purchase orders, invoices, and statements. Many of these documents require signatures for authorization, a situation to be discussed later.

Three initial observations can be drawn (*1–4*). (i) Electronic fund transfers can supplement a large part of today's mail volume. (ii) The major sources of mail, in business and government, are highly concentrated and should be able to economically justify electronic equipment. (iii) Most information to be mailed can be encoded, and the formats of most graphic material are fixed.

Anatomy of Electronic Mail

An electronic mail system is becoming practicable (*5*) today because of (i) the wide availability and decreasing cost of electronic communication channels, originally designed for voice transmission but now enhanced with various wider-band systems; (ii) less expensive electronic control and storage systems that can process information at high speed; and (iii) the availability of technology that can translate information ordinarily handled on paper.

An electronic mail system has three fundamental characteristics—origination, transmission, and printing. Electronic mail can be originated by conventional means (typewriter, handwriting, printing), or the information can be entered on a keyboard device ("keyboarded") capable of either transmitting the information directly or recording the information on some medium (such as paper tape, magnetic card, or magnetic tape) for subsequent transmission. The transmission can be accomplished through a number of channels, such as the direct distance dial (DDD) telephone network, microwave systems, satellites, and packet-switched (*6*) networks. The mail can be printed in any of several ways; by, for instance, impact printers or raster printers using serographic, electrographic, thermal, or ink-jet technologies.

Reasons for Adopting Electronic Mail and Message Systems

Urgency. The delivery time for traditional mail is measured in days. Transmission by facsimile and through most keyboard terminals reduces this transmission by facsimile and through most keyboard terminals reduces this translent, communicating word processors will be increasingly used as electronic mail, particularly when the needs for graphics and requirements for signatures are small in comparison to the urgency for transmission.

Economics. Economics increases the pressure for the development of electronic mail and message systems. One can analyze the comparative cost of transmitting a page using traditional mail methods versus electronic means. The amortization costs of the terminal have been omitted, because there is wide variance of rental costs and selling prices among the various facsimile and communicating keyboard terminals; in addition, some of the devices can be justified for reasons other than their applicability to electronic mail. If we make the assumption that two pages will be mailed in a traditional envelope, the cost is 6½ cents per page. Although the costs of electronic transmission vary, such transmissions are almost immediate as opposed to requiring days. The equivalent transmission by teletype would cost $1.76, with 4000 characters per page. However, given a 300-mile transmission after 11:00 p.m. on the DDD and multiple documents per call, a facsimile terminal with a transmission time of 2 minutes per page can send these two pages more economically at 26 cents each. If we take a communicating terminal as another example, at 1200 bits per second, the transmission cost per document is now 4½ cents per page. One could envision future generations of communicating equipment using magnetic media, such as disks, or high-speed tapes with higher throughput capability and high-speed modems. This would further reduce the cost of transmission and provide a tremendous communications savings for the user, providing his volume can justify the investment in the terminals and systems.

Table 1. Sources and recipients of mail during fiscal year 1968 (*1*). Numbers in parentheses are billions of pieces of mail. For fiscal year 1968, the Postal Service estimate of the total was slightly lower, 82 billion pieces of mail.

Recipient	Source (%)			
	Businesses and other organizations	Households	Government	Total
Business and other organizations	25.0 (21.6)	5.8 (4.9)	1.8 (1.5)	33.4 (28.0)
Households	46.6 (39.1)	14.0 (11.7)	3.8 (3.2)	64.4 (54.0)
Government	1.2 (1.0)	0.4 (0.3)	0.6 (0.5)	2.2 (1.8)
Total	73.6 (61.7)	20.2 (16.9)	6.2 (5.2)	100.0 (83.8)

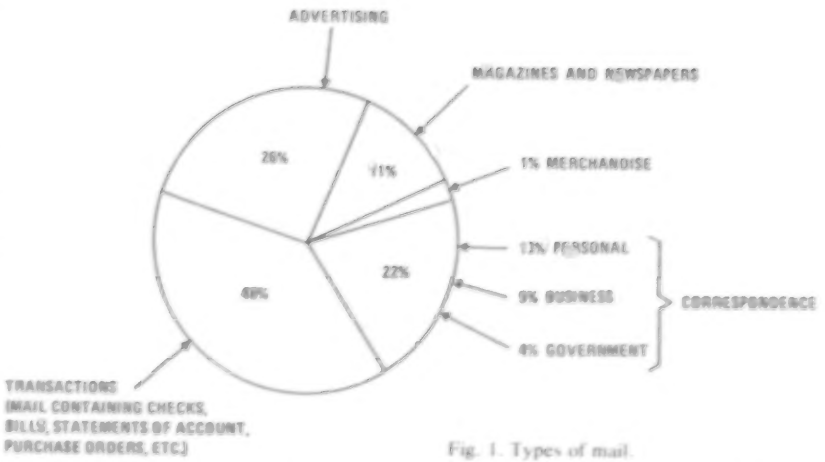


Fig. 1. Types of mail.

Remote and large volume. Electronic transmission of information is particularly useful when the volume is large or the source is remote and the material is already encoded. A good example is a time-shared computer system in which the information is generated in one place but the printing is required elsewhere. The data are communicated electronically to the recipient and are never printed at the origin.

Operational features. In a business environment, the originator conventionally dictates or gives handwritten information to the typist. A draft is typed, returned to the originator, and revised. This cycle can be repeated two or three or many times if the material is complex or particularly important. After the letter is prepared and approved, copies are made and envelopes are addressed and stuffed. Whether by internal mail or the U.S. Postal Service, hours or days elapse until the mail is received.

In a hypothetical electronic mail process, the writing and typing times are the same. However, with a word processor, revisions can be made very quickly, since the letter is recorded, rather than retyped each time.

In addition, the typist can use stored documents from which standard paragraphs, distribution lists, contract clauses, and so forth can be retrieved by using codes (thus avoiding retyping), inserting any portion of the stored material into any part of the newly generated letter. Once the revisions have been made and a final copy has been approved by the sender, the letter can be transmitted in seconds or, alternatively, stored if low cost or other considerations suggest a delayed delivery. The sender of electronic mail is not always its writer; in many cases, one sends mail that someone else wrote. Sometime in the future, the recipient of a document who wishes to pass it on to others, will be able to have access to the electronic file of the sender and direct the transmission of the document to the secondary recipients.

Remote transmission of alphanumeric code is advisable for the bulk reproduction of many copies of a report, procedure, repair manual, or instruction book (among others) when the content has been heavily edited, proofread, or revised, and when reproduction is required at a distance; such documents are ideal candidates for electronic remote printing of the coded information. Although issues of privacy, security, editing, and confidence will arise, it is the way some electronic mail will eventually be handled.

Table 2. Mail findings for 1976 [averages as described in (6)].

Type	Ounces per piece	Pages per piece	Postal cost per:		
			Ounce	Piece	Page
Letters	0.55	1.95	18.2	10.0	5.13
Bulk mail	2.20	7.20	4.4	9.7	1.34

Another form of electronic mail exists when (i) information is handwritten or graphical, (ii) the information on the document has not been encoded electronically (that is, a routine typed memorandum or letter), (iii) when special forms are necessary, or (iv) signatures are required.

There are certain documents which originators want delivered exactly as approved, for example, monthly reports, financial reports, personal memoranda. Most people would not accept a communication of such information in electronic form without being assured that the hard copy will be received exactly as sent with respect to form, format, typographical accuracy, and so forth. In these cases, the electronic transmission of a copy of the document is more appropriate than the electronic transmission of the information on the document.

These documents can be entered into an optical scanning terminal. The electronic image of the document would then be transmitted over either conventional telephone systems or wide-bandwidth channels to a receiving terminal capable of reproducing the scanned image of the original (facsimile or electronic copying). The terminal might be in a mailroom, for subsequent local distribution of the documents, or at the recipient's location for immediate reception. The preparation time is the same as that for conventional documents, but transmission time can be seconds or minutes, unless store-and-forward features (7) are desired.

The handling of approval processes is changed if character-encoded content from nonvisible magnetic recordings is transmitted without inspection by the originator. In addition, a signature is required on certain documents; unless raster scanning is used, it is impossible to electronically transmit the actual signature. However, it is possible to solve this problem and take advantage of the higher efficiency of alphanumeric transmission by using identification codes. Secret codes, similar to access codes on a time-sharing computer, can be assigned to each sender. Upon completion of the document, that identification code can be placed in the terminal, directing that a copy of a signature be printed from raster memory at the remote printer or that an indication of signature authority be typed. In addition, a signature is often unnecessary; many business communications today could eliminate the signature and simply substitute the name of the sender.

Forms such as purchase orders, invoices, and insurance forms have fixed graphic formats. They can be stored in a buffer memory at the receiving end and triggered by a code that is typed into the originating terminal. Some mail can thus be handled by alphanumeric transmission augmented with graphic and alphanumeric printing. Certain graphic transmission will always be necessary and could be handled by strategically located facsimile terminals.

Concern for privacy. The protection of privacy in electronic mail can be ensured in several ways if the system is properly designed. For example, the printer could receive information, not print it, but store it so that only the recipient would have access to it. People will use special or unusual configurations when necessary. For example, during World War II, V-mail was an acceptable way for civilians to communicate with members of the armed forces in remote locations, although privacy was lost and freedom of format was inhibited. Some electronic mail systems may compromise the necessity for a signature or the sealed envelopes in favor of reduced cost or increased speed.

In short, the more impersonal the content and style of the document, the more eligible it is for character-encoded transmission; the more personal or sensitive it is, the more important are the envelope, signature, and actual image.

Applicable Technology

The technology required for electronic mail is becoming sufficiently advanced to break through the cost/performance barrier. While many technological and engineering challenges remain, an awareness of the technology trends convinces us that electronic mail is a realistic possibility. Therefore, it is useful to examine the relevant technology and understand its implications.

Electronic technology. The development of digital electronics is essential to the success of an electronic mail system. The advantages of decreased cost, greater speed, smaller size, and improved reliability of electronic components enable the speed, automation, and cost/performance ratio of electronic mail. Happily, these elements all occur together as circuit elements become smaller and more of them are squeezed onto a single chip. Large-scale integration (LSI) has decreased cost and power consumption by several orders of magnitude while increasing density, reliability, and speed. Electron-beam or x-ray lithography will improve the processes of manufacturing semiconductor chips and thus permit further increases in the number of gates or bits per chip. One can now envision very large solid-state memories at very low cost (*8*).

The effects of these trends are several; for example, it is becoming increasingly attractive to distribute processors (or microprocessors); moreover, localized processors can be made more powerful, permitting a greater range of electronic functions in a terminal at lower costs.

Scanning technology. An input scanner converts the images on a document into raster or bit-by-bit electronic information; scanner resolution is one of the first determinants of image quality. Typical resolutions range from 50 to 500 pixels (picture elements) per centimeter for applications ranging from facsimile to graphic arts. Fortunately, two-dimensional digitized images contain considerable statistical redundancy, and data-compression techniques have been designed to reduce the number of bits to be stored or transmitted.

By the use of digital processors, time-domain or transform-domain encoding techniques have now become practical. Depending on the content and resolution of the scanned document, compression by as much as 20 : 1 is possible. Still, efficiency of encoding can be further increased by increasing the complexity of the encoder and decoder.

Recent scanning advances have permitted the use of laser sources with electromechanical deflection. For the near future, the trend may be toward solid-state scanners such as charge-coupled devices (CCD's) and photodiode arrays because of their higher reliability and lower cost of the subsystems in which they are a part. Vidicons and cathode-ray tubes (CRT's) approaches do not appear to be cost-competitive.

Storage technology. Recording a page of a raster-scanned document typically requires a million bits. Information storage in electronic mail is, therefore, important. During the past 10 years, cost and typical on-line data capacity have changed reciprocally. The cost of magnetic disk storage, at present approximately 10^{-3} cents per bit, is expected to decrease further in the next decade, permitting more economical store-and-forward electronic mail.

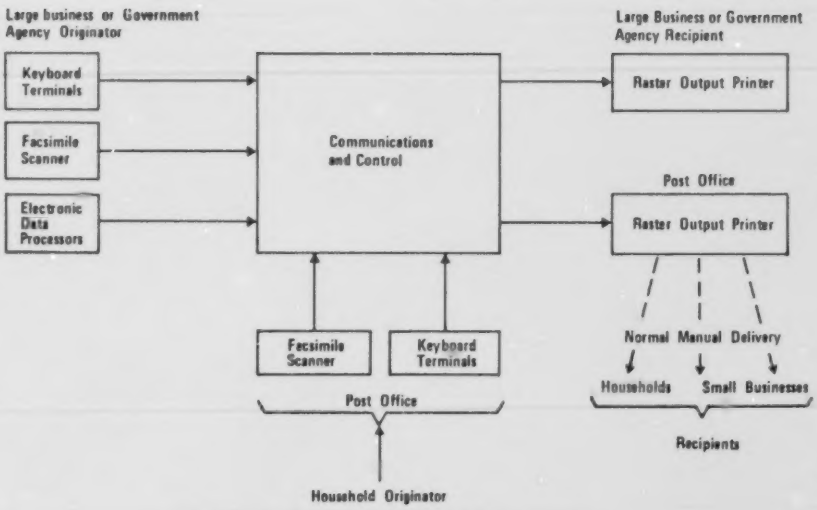


Fig. 2. Hypothetical electronic mail system.

Communications technology. The technology of communications systems is in a period of change and growth. Data rates over conventional telephone lines range from 110 to 9600 bits per second, and sophisticated signal-processing techniques are being employed to overcome the channel-induced distortions at the higher rates. Special wide-band terrestrial and satellite systems are emerging and becoming more economical for electronic mail applications.

Several conceivable configurations could be used with telephone lines or wide-band systems, whether land-based or satellite. Proposals to provide electronic mail systems have been made under a variety of trade and generic names (*9*). It is not the intent to describe here the planned or proposed systems in detail.

Fiber optics promise data transmission with a capacity of 10^9 bits per second. Fibers are expensive now, but their cost may continue to decrease (*10*). Communication satellites will also play an important role in the future of electronic mail (*11*). There are several inherent advantages to a satellite communications system. They include broadcast capability, distance insensitivity, and network capacity on demand. Furthermore, they readily transmit voices, pictures, and data. The use of high-frequency (approximately 10 gigahertz) transmission will avoid conflicts with microwave communications systems as well as permitting the use of small antennas of about 5 meters in diameter, thus allowing earth stations for individual customers and minimizing the need for terrestrial extensions and interconnected systems.

Printing technology. The information contained in a document (after scanning, processing, and transmission) must ultimately be printed. The means used today fall into two major classifications: impact (for example, the mechanical typeface strikes a ribbon transferring ink to paper) and nonimpact (for example, the recorded mark is made by electrostatic, optical, or other noncontact means).

For exclusively alphanumeric, low-speed applications, for which encoded data are captured at the source, single-element character printers have been dominant for years. The print-wheel character printer is becoming more popular today. It will probably continue to be effective because of its inherent print quality, speed, and costs.

Printing of high-resolution raster-scanned images and high-speed alphanumeric information is possible today only with nonimpact printers. Printing resolutions vary with the application and are typically between 50 and 200 pixels per centimeter. (i) Ink-jets provide high-speed serial printing, with print quality that appears to be acceptable for many documents. Further improvements in both print quality and speed are possible. (ii) Laser exposure and xerographic recording are currently used for facsimile and computer printing. They can also be used for combined printing of text and images at moderate speeds. (iii) Electro-

graphics can print at high speeds, but it suffers from several limitations. These include the requirement for expensive special paper and somewhat lower print quality than is now available from impact printers. (iv) Matrix impact printers are faster but produce prints of somewhat lower quality than those that print one character at a time. (v) The electroresistive recording process is cost-effective for low-speed facsimile applications, but the special paper becomes too expensive for large-volume applications. (vi) Thermal printing is still used occasionally, but low resolution (approximately 25 to 50 lines per centimeter), the requirement for special paper, and the lack of archival capabilities appear to limit its future potential.

The high-quality serial printer will probably continue to use impact character printers. The performance of ink-jet printers will continue to improve, and their acceptance may broaden.

Character-Encoded versus Scanned Information

Systems must be able to accommodate both encoded characters and graphics. Examining the efficiency of character-encoded transmission versus raster-scanned transmission quickly shows why it is important to use character-encoded transmission and record keystrokes whenever possible. In today's business facsimile with a resolution of 96 by 96 lines per inch, an 8½-by-11-inch page will contain approximately 860,000 pixels. If this page is transmitted in digital form (black and white only), 860,000 bits of information need to be transmitted.

With today's technology of data compression, with removal of the redundancy resulting from white spaces in a document as well as image redundancy in both the vertical and horizontal directions, it is possible to encode the information in a more efficient manner. Commercially available systems are capable of about 7 : 1 compression ratios; thus, the technology can now reduce the original 860,000 bits to approximately 123,000 bits. If we assume a transmission rate of 9600 bits per second, the transmission time for a typical page is 12 seconds.

In comparison, consider the time required for a character-encoded system. One page will contain approximately 3,000 characters (or 24,000 bits). The time required for transmission at the rate of 9600 bit/sec is 2½ seconds, approximately five times faster than transmitting the image. This increased speed can result in decreased costs, but the form and format will probably be lost. Whether the system uses the traditional telephone network or a wider-bandwidth system, it will probably be billed in such a way that high speed transmission is desirable. Given the efficiency of character-encoded transmission it is easily understood why a system should record keystrokes whenever possible. Also, the system should be able to handle graphics, for example, schematics, diagrams, written words, and typed material from ordinary typewriters, printers, and secondary sources. To accomplish this, multifunction electronic mail terminals will probably be developed, as both alphanumeric and image information must be handled, separately or together.

An Electronic Mail System

We now can sketch a hypothetical electronic mail system, as shown in Fig. 2. Large businesses or government agencies have communicating word processors, facsimile scanners, and data processing centers can have access to the electronic mail system. Large agencies receiving the messages will have raster output printers capable of handling both encoded information and images. Thus, between businesses and government agencies, an electronic mail system is possible.

However, one cannot overlook the small businessman and the household. It is unlikely that, during the next few years, the cost of a printing terminal will fall so low that they will be available to all private individuals. However, it is possible that an individual post office would have an output printer that would permit it to receive messages from businesses and government agencies, as well as from other postal locations. Such information could then be normally mailed to households and small businesses. Again, the post office is able to accept messages from a household from either a document to be scanned or through the telephone. This would, in turn, be typed on a communicating device. The central electronic mail system will provide electronic message switching, store-and-forward and accounting data, such as traffic volumes and billing.

Conclusion

We are in the midst of the evolution to the electronic communication of information, which will replace some traditional mail. To this end the process of adopting emerging technologies for electronic transmission of the image of documents to be mailed will continue.

The technologies are in hand and can be applied either by the conventional telephone system or the emerging wide bandwidth of the satellite system. Facsimile terminals for image transmission are available and increasing in number. Terminals for printing alphanumeric code are present and becoming more widespread. In the future there will probably be devices that can handle character and images simultaneously.

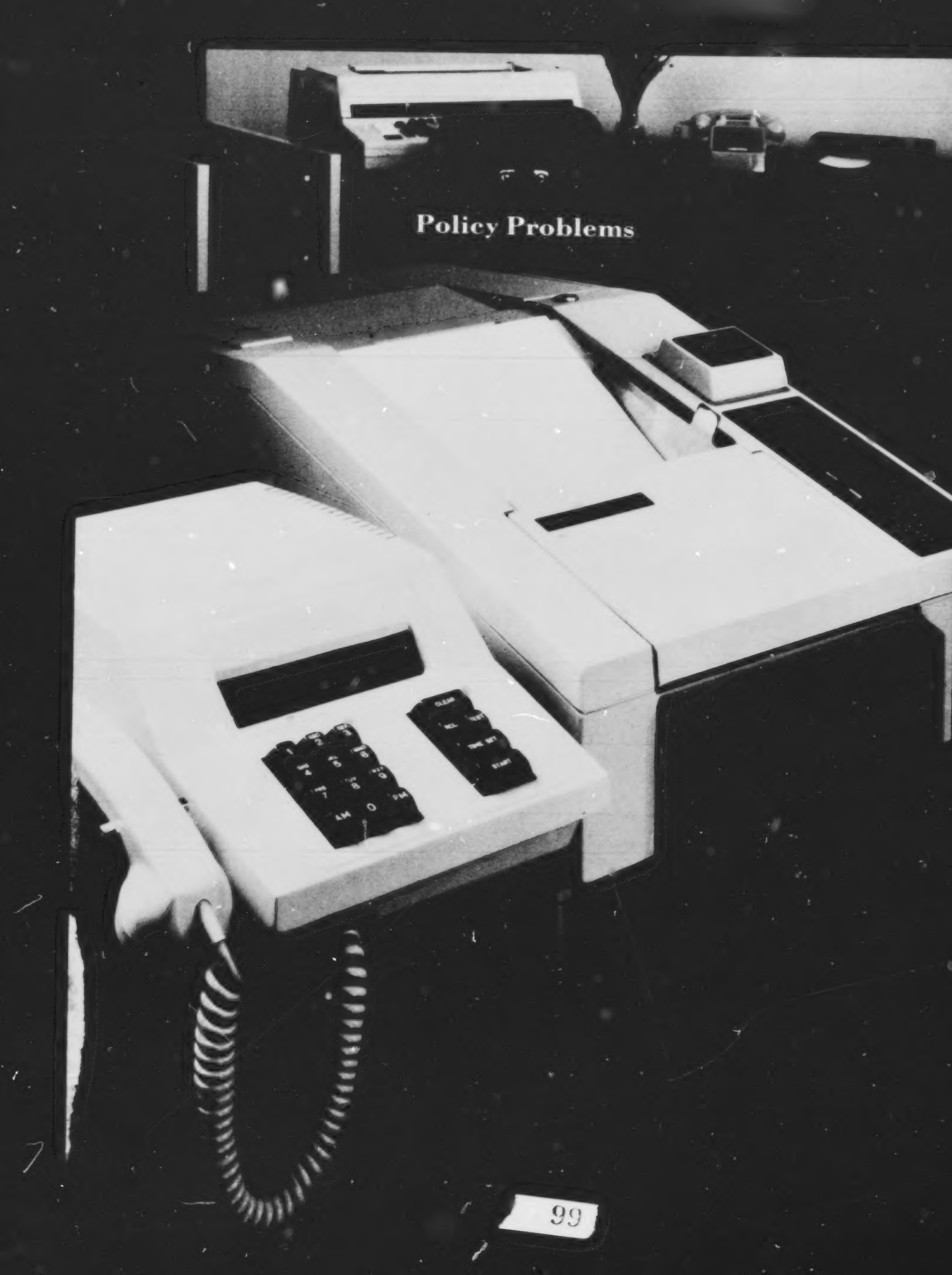
Since the sociopolitical and economic issues are complex, it is beyond the intent of this article to address the question of who should own or operate the electronic mail system. The evolution is inevitable for the reasons I have discussed. But, in addition, an indirect force will stimulate the growth of electronic mail; physically transporting tons of paper on a daily basis is an inefficient use of our resources if the information can be transmitted electronically. The impact of electronic mail will be first on large companies and large government agencies; it will gradually be felt by individuals and small businesses.

References and Notes

1. President's Commission on Postal Organization, *Toward Postal Excellence* (Government Printing Office, Washington, D.C., 1968).
2. C. L. Jackson, *MIT Ind. Liaison Program Rep. CSR TR-73-2* (1973).
3. *Facsimile and Electronic Mail* (International Resource Development, Inc., New Canaan, Conn., 1976); *Communicating Word Processors* (International Resource Development, Inc., New Canaan, Conn., 1976); T. Johnson, *New Sci.* **7**, 351 (1976); A. Purchase, *Stanford Res. Inst. Guidelines Rep. 1001* (1976); R. W. Hough, C. Fratessa, V. Holley, A. H. Samuel, L. S. Wells, "A study of trends in the demand for information transfer" (Stanford Research Institute, Menlo Park, Calif., 1970).
4. Satellite Business System, a report to the Facsimile Suppliers Conference, Washington, D.C., 2 December 1976.
5. Quantum Science Corporation Conference, New York, 8 February 1977.
6. A computer-controlled communication network that transmits information in fixed groups, for example, 1024 bits per packet.
7. A technique whereby data are stored in a buffer, either locally or remotely, for subsequent transmission.
8. J. Rajchman, *Science* **195**, 1223 (1977).
9. *Commun. News* **13** (No. 6), 27 (1976); *Geyer Industry Letter* **1** (No. 20), 1 (1976); A. Vezza, *MIT Ind. Liaison Program Rep.* (1975).
10. S. E. Miller, *Science* **195**, 1211 (1977).
11. B. Edelson and L. Pollack, *ibid.*, p. 1125.

12. I thank John Tambert of Xerox Office Systems Division for staff help in developing the content, researching the references, and accumulating knowledge and background from many people.

Xerox Corporation, Dallas, Texas →



Policy Problems

The Convergence of Computing and Telecommunications Systems

David Farber and Paul Baran

A major change in computer communications is taking place. As the cost of communications and computation declines, the number of new services that can be offered will increase and a large number of new industries can be expected to emerge. A battle between the potential suppliers is now under way.

This article considers: (i) the nature of the new services, (ii) their importance to the economy, and (iii) the arguments for regulatory control as proposed by the major potential competing suppliers. The basis for seeking regulation hinges primarily on the use of a communication component provided by a natural monopoly that in the past has been highly regulated. Rapid changes in the technological base make it difficult, if not impossible, to resolve the differences between the three major competing suppliers [American Telephone & Telegraph Company (AT&T), International Business Machines Corporation (IBM), and the smaller computer companies] within the existing communications regulatory framework. Considering computation and communications as distinctly separable functions does not provide the basis to resolve boundary disputes rapidly when regulatory issues are involved. This article examines the underlying technical changes, reflected in the conceptual changes, which now make the regulators' tasks difficult. It suggests that part of the present turmoil and uncertainty is a result of the growing use of digital processes within the communications industry itself—an inevitable evolution of the new technology. An appendix describes the evolution of data transmission technology.

We do not intend to propose solutions to the very difficult problems that this article raises. We do not know of any completely workable solutions. Our purpose is rather to sensitize the reader to the issues and problems that our country will face now and in the future.

Mr. Farber is associate professor of information and computer science and electrical engineering, University of California, Irvine 92717. Mr. Baran is president of Cabledata Associates, 701 Welsh Road, Palo Alto, California 94304.

The Nature of the New Services

This article was written by two authors 500 miles apart, using a computer communication system. Each writes on the same sheet of "paper." Each amplifies, modifies, and clarifies the words of the other. Misspellings are corrected, and the order of phrases is changed. Each author takes his turn tidying up the modified manuscript by pressing a few buttons and giving a few commands. This takes care of the nitty gritty in paper writing, gobbling up the blank spaces and even justifying the right margin. Discussions on content and outlook take place with the use of the same computer communications-based message systems. The method used to produce the article you are now reading is not tomorrow. It is today.

Tomorrow, computer communications systems will be the rule for remote collaboration, like the above example. As computer communications systems become more powerful, more humane, more forgiving, and above all, cheaper, they will become ubiquitous. The obvious next question is, "Why bother to retype this manuscript manually, set type, and print this journal on paper?" Once the words in this article have been entered into the computer system, they are in fact "published" if a copy has been stored in a publicly accessible memory portion of the computer communications system. Think of the paper wasted in conventional periodical publications when only a small portion will ever be read.

The cost component of the computer communications cost-demand curve declines each year. Yet improvements of literally several orders of magnitude are possible on the basis of what is known already. Accelerated expansion of demand is a certainty. With increased use of computer communications, changes in our institutions will follow. The example given for publishing is only suggestive of the range of change. The immediate reaction of those in publishing to such change is to feel threatened. Institutions rarely take kindly to the prospect of obsolescence of their present justification.

Importance to the Economy

Our simple illustration of the potential change in publishing could, and most likely will in time, spread across the entire economy. Even today we are an information-processing society. More gross national product is developed in the United States today in activities in the information sector than in the production of tangible goods.

We have chosen a dramatic application of computer communications solely for illustration. More important are the mundane applications: automated hotel reservations, credit checking, real-time financial transactions, access to insurance and medical records, general information retrieval, and real-time inventory control in business. A wide range of services such as these and beyond is coming. Automation of the office is almost a certainty. The only questions are "when," "how," and "by whom?" Electronic transmission of much of the current first-class mail similarly will follow.

Change will be felt widely throughout the economy and, in a fundamental manner, by many institutions. Such change is generally perceived as crisis and rarely accepted with grace. Most people believe that change is something best absorbed by others. The interim period will be a period of inevitable discomfort to those organizations and professions that are affected.

The Argument for Regulation

History of potential suppliers in this information economy. Although computer communications promises to provide huge new markets for potential suppliers, there is uncertainty about who they will be. It is not simply a choice of AT&T, IBM, or a diversity of smaller computer companies. Rather, the question is which supplier will capture the largest, most profitable slice of the pie, and with the minimum degree of the unpleasantness of competition.

Each supplier arrives at its claim of legitimacy to supply the new markets by a different route. The telephone industry supplies communications and is now a regulated monopoly. (The most economical way of providing basic communication channels is on a common shared-resource basis—it is almost unarguably a "natural monopoly.") The IBM claim in this new industry is well-founded. It will provide a significant share of the new market's hardware. It is cash-rich and thus unique in the computer industry in having the financial resources to make the needed capital investments. And it is the winner in the proof-by-adversary test in the free economy. It has been tried and has won in the free marketplace. The third contender, the remainder of the computer industry, feels that IBM has won unfairly and seeks to handicap the giant. (There are few computer companies today that have not brought suit against IBM for alleged violations of anti-competitive statutes.) These smaller companies feel that IBM may exercise monopolistic power in a sector in part deemed not to be a natural monopoly and thus not regulated. This may all seem strange in a free economy. But the regulatory and institutional constraints are so treacherous that the outcome of the debate may hinge more on political adroitness than on market acceptance.

The disintegration of the boundary between suppliers. The basic premise that communications functions and computational functions are separable has been rendered meaningless by a basic and pervasive change in technology. The separate "atoms" of communications and computers form a new "molecule" with properties different from those of their atomic components.

Consider the recent case in which AT&T wished to offer its subscribers the sophisticated model 40 communications terminals made by its wholly owned subsidiary, Teletype Corporation. A series of delays ensued as competitors argued that AT&T was entering the processing business; AT&T said "nonsense, it is just in the communications business." The Common Carrier Bureau of the Federal Communications Commission (FCC) decided against AT&T's position. After a short delay, the commissioners overturned the decision of their own bureau. Indecisiveness and on-again, off-again decisions at this boundary line are common. There is no longer an unambiguous dividing line upon which to reach decisions. The lawyers involved in resolving these arguments are assured of an employment stability lacking in most other sectors of the economy, as the endless suits at the interface between suppliers and the government seem to be careers rather than occasions.

The cause of the disintegration. The regulatory paradigm assumes that communications is communications, and computing is computing. Unlike oil and water, they now mix well—too well to treat separately, once combined. They simply are no longer separable into orthogonal components subject to different legal treatment. Separating communications and computational processes is becoming an almost nonsense activity, like unscrambling an omelet. It is a task better left to philosophers than to lawyers on regulatory commissions.

This has not always been the case. The basic changes have crept upon us gradually. The forces melding these two previously separable domains are several. At a gross level, there is the economics of distributed processing. It may be cheaper in the future to provide independent computing power for those who need it, at their terminal, office, or home. But just having one's own computer is in general not adequate. There is the occasional or possibly frequent need for shared information where, irrespective of cost, access to remote shared data in real-time is mandatory. But, most important, the basic nature of the communications process is that digital bits provide the natural common form for all signal transmission and processing. The digital wave form becomes the universal package with which to send voice, data, facsimile, or anything else. Why this is so, and how it got that way, provides insight into the evolving technology and the increasing difficulty that will be felt by those who try to reach decisions on the premise that communications are separable in computer communications networks. What are the alternatives available for solving our dilemma?

Why aren't microcomputers a perfect substitute for large interconnected computers? There is a school of thought that argues that computer communications is a chimera. There is no need to use the resources of AT&T or IBM. According to this viewpoint, computers are going to be so small and cheap in the future that we won't need to use communications lines to talk with big computers. Is the low-cost mini-, or micro-, or picocomputer the answer? Powerful computers have become cheap and will get cheaper. The explosion of the computer hobby market during the last year and a half has been remarkable. There are now 150 specialty stores selling low-cost computers in hobby kit form. More computer development work is being done by hobbyists today than by computer science researchers funded by public agencies. The hobbyists are not all amateurs. There are some whose vocation is also their avocation. They are intelligent, dedicated persons no longer blocked from entry into computer research by the cost of the equipment. And there are many more now at work (or play?) than the government could possibly support.

The reasons why access to communications capability will always be a necessary adjunct to the small computer are basically unchanged:

1) Whenever it is necessary to make infrequent use of a highly specialized resource, it is cheaper to share (why build a machine to make nails when you can buy the few you need at the corner hardware store?).

2) The output of one data processing task is usually the input for another process.

3) By agglomerating computing power as a result of the interconnection of a number of previously independent systems, applications that are currently not feasible in the data processing area can be handled. These include (i) systems which provide high availability, (ii) systems which can grow modularly as their environments grow (the problem of what to do with a system that supports N users when the $N+1$ user arrives), and (iii) systems which match the physical organization of computing power with the corporate organization, thus relieving organizational strains and stresses.

Distributed processing technology and its fulfillment. All of the above applications can be satisfied by distributed processing, loosely defined as the use of a multiplicity of computational devices for a single processing task. Distributed processing implies that the units that compose the distributed computer can freely intercommunicate. If the distribution is other than local, then remote computer communications facilities are required.

Data-processing begets data-processing. Consider the following example suggesting the nature of distributed processing. A cash register in a grocery, really a counter computer, computes a customer's bill quickly. But it also generates the information needed by the store to order fresh turkeys today for delivery tomorrow. This same information can be used by a hypothetical turkey farmer to determine instantly whether there will be a shortage or a surplus of his product, and

whether to hold out for a better price or sell his turkeys now. This information has economic value to his procurement decision. The key feedback signal of a free economy in the past has been the flow of money. More rapid signals can be obtained as an inexpensive by-product of the flow of information from other processes.

Economic examples are not the only use for computer networking and distributed processing. Equally, or more significant, are those that deal with human communications. Just as in the preparation of this article we were able to interchange ideas and manuscripts over the computer communications media interactively, so users of the low-cost computer systems of the future will want to intercommunicate for ideas, pleasure, and business. One has only to look at the meteoric growth of citizen's band radio to realize the new opportunities for communications between people in our society. Microcomputers with a near-zero cost making efficient use of a shared communications resource can open up entirely new forms of human communications.

The digital channel—the evolving choice for communications transmission. Before considering this and other issues in detail, it will be helpful to review some of the characteristics of the digital channel and the reasons why it is becoming desirable even for the communication of basically analog signals. An analog signal is a voltage sent over the usual telephone line representative of (or an analog of) the pressure wave form of the human voice. Why should such a conceptually simple wave form be translated into a series of bits? The reason is that, once converted to digital format, it is possible for the wave form to tolerate a tremendous amount of distortion of its wave shape (the signal) and still be capable of simple reconstruction into a neat distortion-free series of pulses. This allows greater economy in the packing of information in a communications circuit and allows the use of lower quality (and cost) channels. Once in digital format, pulses that have been lost or converted into the wrong symbols are recoverable.

It is relatively easy to carry out many rapid calculations upon the data stream to detect and correct errors. (The price is a small cost for a few extra bits to serve as error detectors and correctors.) The ultimate in such distortion removal is cryptography, where a bit stream is purposely modified by a known-in-advance bit stream (called the key). One "subtracts" one binary wave form from the other and reconstructs the initial wave shape without distortion of the transmitted intelligence. An unauthorized observer sees only a binary stream with a seemingly random chance of either being a "1" or a "0," where "1's" and "0's" are the names of the two states possible in the digital channel. Affordable secure transmission mandates digital transmission. The digital stream simplifies the task of transmission design. Here again, if the signal does not become too distorted, below 50 percent of its initial value, it can always be reconstructed without errors. Digital transmissions, unlike analog transmission, which requires critical adjustments merely to minimize distortion, is thus the ideal technique.

Switching. Communication networks require both transmission and switching. Switching is the process of connecting any network user to any other network. Switching, by its nature, is a digital process. In the 1930's, an operator plugged one circuit into the next of a tandem chain of connections through telephone switching offices. This is a digital process as each switch is either open or closed, a neat binary process. The decision process of laying out switches and relays for telephony was among the earliest practical uses for Boolean or binary logic. The foundation of the logical design process in computer design owes much to the telephone industry. Moreover, the integrated circuit devices which make real-time computer applications feasible are all offsprings of the transistor developed at the Bell Telephone laboratories. The telephone dial itself is a digital device. A series of pulses from the dial acts to set switches.

As computer technology matured to the point where its products were realizable as well as elegant and complex, the quaint electromechanical computers that formed the telephone switching network of the past were replaced by computer circuit logic. This evolution continues as digital transmission is now being combined with digital switching. The two separate processes in time are melding into a single integrated process. Bits going into one end of the network are thus beginning to be treated almost the same way as a bit stream in a computer, albeit with less processing.

What Is the Regulatory Issue?

Who should be the future legitimate supplier of the digital system? As the telephone technology slowly evolves toward an almost all-digital system in the far future, it will or can develop the technology needed to provide many of the new computer communications-based services. For example, AT&T has already developed devices for credit card transactions and sophisticated computer input-output terminals through its Teletype Corporation.

Equally comfortable in the domain of processing and perhaps even transmission is the highly competitive computer industry. During the last decade, the FCC encouraged competition in transmission for other than voice telephone services. New transmission companies and independent "value-added carriers" were authorized. The value-added carriers essentially lease telephone channels from the communications carriers, add processing, and use the leased lines so effectively that they are able to sell the joint product for a profit over and above the cost of the basic regulated channel.

Today, much of the in-fighting involves legal and regulatory actions seeking to block new service offerings and capabilities. It is, in part, an attempt to prevent the opponents from establishing beachheads by precedence. And, in part, it is a way of adding uncertainty to the industry to discourage capital investment. Any new computer communications service offered today by a potential supplier is certain to be opposed by the other major contenders. IBM's plan for a communications satellite network for major business users, which bypasses AT&T Long Lines, faces opposition from both AT&T and computer companies fearful of IBM's assumed long-range intent. AT&T's offering of sophisticated terminals is met with countervailing opposition from the suppliers of computer terminals. The outcome of the competition between the suppliers will be decided, in part, by the staying power, legal resources, and political muscle each allocates to the fray. Little attention is paid to the "public interest." The term defies definition. Is the public interest the interest of the cross-subsidized residential telephone user? Is it the interest of a business which faces a reduced communications bill? Is public interest to be viewed primarily in the short term, irrespective of long-term damage to existing institutions in achieving immediate savings? There is no simple calculus to reach decisions. At present, there is a struggle between those who point out the advantages of competition versus those who hold that better use of limited capital results by virtue of monopoly. It is unclear which view will prevail. Until it becomes clear, resources will be wasted

and future economically beneficial services to the nation will be delayed simply by the inability of industries to cooperate.

Summary

Since we do not know of any completely workable solutions to the problems we have considered, we shall settle for a few general observations:

1) The public availability of socially useful computer communications services is and has been held back by legal battles that are now under way between the potential suppliers.

2) No simple resolution of these issues in the near future seems likely in view of the past conceptual separation of computers and communications doctrines.

3) The current policy is to determine whether the nation shall or shall not have certain computer communications services, by the adversary process. In this process, often only the voices of the loudest adversary suppliers are heard.

4) Although there can be no certainty that better alternatives cannot be devised, we believe that such a possibility assumes a higher probability if the key actors come from the technical community sectors more representative of the future consumers.

5) If we are to have the new services that are possible, we need an approach that makes better use of the technologists' dreams and goals rather than have future prospects excessively bound by lawyers paid to preserve the interests of their clients, irrespective of any secondary consequences.

6) We cannot be sanguine about this possibility, as technological statesmanship is too easily corrupted by the same forces that have placed us in this predicament. Furthermore, even if not corrupted, beneficial cooperation can too readily be regarded as simply collusion.

7) Although we do not have any clear answer, we do know that present approaches are not taking us where we want to go very rapidly and that alternative approaches should at least be considered.

Appendix: Advances in Digital Computer Communications Technology

We consider here the technical evolution of digital communications.

Modulation. The first digital transmission circuit was the electrical telegraph, predating the voice telephone by 35 years. Although local telephone lines can be and are used for simple on-off direct-current (d-c) signals, the need to transmit signals beyond the limited distances where a d-c path exists led to the use of tone or alternating-current (a-c) transmission.

The device that converts the binary signal into an a-c signal for transmission is called a "modulator." The device that converts the a-c tone back into a properly shaped digital signal is called a "demodulator." The generally used arrangement, called a "modem" (MOdulator/DEModulator), combines both functions.

The first of the present-day-type modems was developed in the 1950's primarily by researchers at the Lincoln Laboratories of the Massachusetts Institute of Technology for transmitting digital signals between air defense sites over analog telephone lines. The efficacy of these devices has improved steadily since then as a result of improvements from a number of organizations. Today we are reaching perhaps 50 percent of the theoretical transmission rate of the conventional telephone line. Modem costs have been declining steadily as the signal processing within the modem is itself converted to digital processing. This conversion is carried out effectively with the use of large-scale integration semiconductor chip circuits. The telephone plant in existence today provides a widespread access channel with low entry cost for many of today's computer applications.

Bandwidth availability. The figure of merit for a modem is the number of bits per second of digital information that can be transmitted over a conventional telephone channel. Data rates on the order of 300 bits per second (about 30 characters per second) simultaneously in both directions are readily achieved at low cost without electrical connection by means of acoustic coupling over any telephone line. Data rates as high as 9600 bits (one way) (about 960 characters per second) are achievable over most (but not all) voice-grade circuits with relatively expensive state-of-the-art technology.

Transmission carriers lease full-time circuits with greater bandwidths than voice circuits. Millions of bits per second can be transmitted over television bandwidth channels. Naturally, the tariffs for such arrangements are expensive, and the applications for such channels limited.

Multiplexing. Even though broad bandwidths are technically feasible, most applications can be satisfied by low data rates. A local telephone call or circuit rental has historically been inexpensive, whereas long-distance circuits have been expensive. The cost of a short-period terminal connection to a computer on the other side of the country, an expensive business, is equal to a long-distance voice call, albeit few bits may be exchanged.

The answer, of course, is combining many users' signals to share a single expensive long-distance channel. Many telegraph, and later, data signals were combined together on a single line. This process is called multiplexing. One form is called "frequency multiplexing." Here each user is assigned a different frequency, and the individual frequencies are separated at the receiving end by tuned filters. Alternatively, in time division multiplexing each user's bit streams are interleafed sequentially. Fundamental to the notion of multiplexing is the fact that each data user will not be transmitting most of the time. The fastest typist at a keyboard generates only a few tens of bits per second. For the hunt-and-peck typist, the result is only a few bits per second. The telephone channel can readily transmit 4800 bits per second, providing a potential for sharing.

Statistical multiplexing. When there is a cluster of data users who wish to send signals to a distant computer, a single expensive circuit can be used. To pack more users onto the line means taking advantage of the fact that, at any one instant of time, only a portion of the possible users would want to transmit. Moreover, to make the averaging even more effective, advantage can be taken of readily available digital storage to hold data and thus smooth out the peaks. "Time buffering" increases the efficiency of the multiplexing process. So devices were developed that combined signals from many users and elegantly packed them together for transmission over a single line.

Spatial multiplexing or packet switching. Multiplexing is highly effective if all users who want to intercommunicate with one another are in only two locations. If the potential users are spread all over the map, the same statistics can be utilized, although it becomes necessary to identify uniquely where the bits came from and where they are to go, and to add a little housekeeping data to detect damaged packets in transit. To minimize the processing equipment, the bits are formed into consistently organized packages called packets. Such packet switching appears on the ascendancy. Most widespread computer communications networks to be built in the future probably will use packet switching, at least in part.

Packet switching networks are now

being built in the United States, Canada, Europe, and Japan. Interconnection between these networks is planned. Packet switching has several advantages: (i) a very robust structure can be built that readily permits the building of systems whose operation is much more reliable than its elements (communications lines and packet switching centers); (ii) it provides the highest degree of statistical averaging to make most effective use of the basic resource; (iii) the standard format package simplifies full effective interconnection between completely different computer systems and terminals; and (iv) no better alternative is in sight for most user-to-computer and most computer network applications.

Future directions for digital networks. Packet switching, although important, is by no means the end of the development for future computer communications systems. The telephone plant itself is moving on to a mostly digital structure in order to achieve a number of economic advantages. In the process it looks more like a computer communications network but with tremendously greater data-handling capacity. One might argue that the two systems are really the same. However, the minor differences between the two systems are certainly of major significance. An argument can even be raised that the conventional telephone set is nothing more than a "terminal" which generates and accepts commands and which receives and generates data to other such instruments, with the "data" being digitalized voices. A single digital voice channel in today's digital telephone systems carries 64,000 bits per second without a conventional modem. Clearly this is a tremendous capacity. There are future needs that could well use some of these high-capacity trunks of the telephone plant. There is a need in the distributed processing area for large bandwidths. In the case of intercomputer file transfers or access, bandwidths on the order of two to three megabits are appropriate whereas for the interchange of high-fidelity pictorial information 10 to 50 megabits are necessary.

Broad bandwidths are not needed everywhere. The same situation prevails in the local versus long-distance distribution of data as in the local versus toll distribution of voice communication; namely, the cost of local circuits is much less than that of toll circuits. In the case of local distribution of high-speed data, new systems such as the Pierce loop developed at Bell Laboratories, the Irvine ring designed at the University of California, Irvine, and the Ethernet produced by the Xerox Palo Alto Research Center have been proposed and constructed to attack the problem. These systems are marginally applicable to the distribution of such capabilities over large distances. The interfacing of such local data networks to national and international networks represents one of the present edges of computer communications technology.

The Information Economy and Public Policy

Manley R. Irwin and Steven C. Johnson

If a single word summarizes U.S. telecommunication policy, that word is conflict. Telephone markets are vulnerable to competition from specialized carriers, the telephone industry is challenged by domestic satellites, telephone equipment is experiencing increased competition from outside manufacturers; the computer and communications industries are engaged in skirmishes for the lucrative market of data processing.

And this is only the beginning. Both state agencies and the FCC have squared off in a quarrel over their respective regulatory jurisdictions; controversy has erupted as to the proper policy mix between competition and regulation; both government and private antitrust suits have challenged the structure of the telephone carriers; and legislation has been introduced declaring a moratorium on competition in telecommunication.

In this article we attempt to provide the framework in our search for answers to questions concerning the cause of the conflict, the forces that contribute to it, the policy alternatives that are open for future consideration, and their effects on service markets.

U.S. Telecommunications

Conflict surfaces when two concepts, two ideas, or two philosophies find themselves juxtaposed. The telephone carrier industry, its services, its equipment, its organization—indeed the premise of regulation—is today subject to scrutiny and reappraisal. What has been the basis of the U.S. telecommunications industry for the past decades? The answer must begin with an examination of the telephone industry.

Service. Telephone companies have traditionally regarded their mandate as rendering a universal service to all subscribers (*1*, p. 4). Such universality, plain old telephone service, presumed carrier ownership of transmission, switching, terminals, and local loop plant. Local exchange services were priced at flat rates within specific areas and tariffs between customer classes rested on the concept of service value (*1*). Long distance tariffs were determined by the variables of time and distance.

Telephone companies were assigned exclusive service areas. Costs were aggregated or spread among the various types of equipment, vintages of plant, and density of population areas. Inevitably, subscribers in some areas subsidized customers in others.

The industry was capital-intensive. The investment required to establish a large telephone system was deemed to erect prohibitive barriers to entry. The industry was also said to enjoy economies of scale or size that gave monopoly a social advantage over competition (*2*). The results saw the industry evolve into a private monopoly—a development that laid the foundation for government regulation.

Dr. Irwin is a professor of economics in the Whittemore School of Business and Economics, University of New Hampshire, Durham 03824. Mr. Johnson was formerly with the Federal Communications Commission, and is now a staff economist with the National Research Council, Washington, D.C. 20418.

Equipment. The concept of universal telephone service carried with it a corollary. Subscribers leased service rather than equipment (*3*). The carrier assumed full responsibility for a total, end-to-end concept, a concept that mandated control over equipment, its installation, its maintenance, its obsolesence, and, ultimately, its retirement (*2*). Subscribers, in fact, were precluded from attaching their own equipment to telephone lines under penalty of service loss or withdrawal. By definition, competition in supplying equipment to the telephone subscriber was deemed contrary to the public interest.

Organization. The telephone holding company became the organizational norm for the industry. The holding company did not render telephone service per se, but rather controlled firms that either supplied service or manufactured equipment. Holding company control of both manufacturer (seller of equipment) and the utility (the buyer of equipment) evolved into vertical integration (*2*). That ownership bond became the generic model for the structure of the entire telephone industry.

As the industry leader, the American Telephone & Telegraph Company (AT&T) controls 22 operating companies, a manufacturer (Western Electric), and Bell Telephone Laboratories. Bell's vertical integration eventually was replicated by non-Bell or independent telephone companies, including General Telephone, United Telecommunications, Continental, Mid-Continent, and others. Today more than 90 percent of the U.S. telephone service is rendered by companies owned or controlled through holding company ownership.

Holding company control was not merely paramount, it was functional as well. Research and development (R & D) laboratories designed equipment, manufacturers fabricated hardware, and utilities bought and incorporated equipment into their investment rate base. Central administrative control—the holding company—coordinated the components of finance, engineering, purchase, and service.

The results of the telephone holding company have been impressive by most standards. Telephone investment in the United States today exceeds $80 billion. Annual revenues exceed $35 billion. The industry's annual construction budget approaches $12 billion. Some $1 billion is allocated to R & D alone. The equipment market exceeds some $7 billion annually (*4*).

The Bell System accounts for most of the nation's telephone equipment and telephone service. Bell provides some 80 percent of all local exchange service and 90 percent of toll message telephone service. The rest is rendered by 1600 independent telephone companies, serving 50 percent of the geographical area of the United States. More than 94 percent of the homes in the country enjoy telephone service (*5*).

Under holding company control, each telephone system moved in the direction of self-sufficiency. Western Electric, Bell's manufacturing subsidiary, supplied the bulk of telephone hardware to the Bell operating companies. The manufacturing affiliate of General Telephone did likewise, a pattern repeated by other independent holding companies. In short, the holding company vertical integration and the systems concept evolved into closed, self-contained entities.

Challenge to U.S. Telecommunications

Service. The traditional triad of service, equipment, and organization has found itself challenged. Today, instead of universality, the telephone industry is experiencing competition, especially in private lease submarkets and in intercity facilities. Terrestrial microwave, digital satellites, and digital and packet switching obviously broaden the notion of service beyond the ordinary analog telephone. New subscribers, particularly those vesting interest in data communications, demand service that is tailored to their individual requirements. A number of firms have responded to the data communication market, a form of rivalry that has proved disconcerting to incumbent telephone carriers. In short, competition for intercity services represents a conflict between telephone carriers and new entrants in specialized services.

Equipment. Competition in telecommunications equipment represents a second area of contention. No longer are telephone handsets the only apparatus attached to the switched network. Today computers are interconnected to remote data bases, data terminals are linked by various networks, and point-of-sale devices are pervasive in their application. Some predict that tomorrow's office will tie copiers, dictaphones, word processing equipment, electronic files, and microprocessors to telephone lines (*6*). Thus, the telephone handset no longer stands alone, but is rather one part of a broadening spectrum of telecommunication hardware, a spectrum that is evolving into a wide range of information terminals.

The proliferation of telecommunications equipment is driven by the needs and requirements of the business user. Acting under the pressure of the marketplace, the business subscriber seeks equipment that yields lower costs, greater flexibility, and increased productivity. That search, sprung loose by regulatory action in the late 1960's, has created market opportunities for equipment suppliers, manufacturers outside the holding company family. These manufacturers sell private branch exchanges, data modems, key telephone systems, and automatic call distributors—to mention but a few.

The business subscriber has accordingly been exposed to a broad selection of equipment. In fact, some subscribers prefer hardware not manufactured by the telephone carrier suppliers. Once the user avails himself of that choice, however, the investment rate base of the carrier is eroded. Telephone utilities strenuously oppose rate base diminution on grounds that revenues lost in one market must be compensated through higher rates from other markets (*1*, p. 21). Inevitably equipment interconnection represents a second dimension of controversy in the telecommunications industry.

Organization. A third area of conflict is organizational. Given vertical ownership of buyer and seller, telephone carriers traditionally buy the bulk of their hardware from in-house suppliers, a practice that has limited the market opportunities for nonintegrated firms. But these practices, long unchallenged, have suddenly been criticized as independent manufacturers seek, solicit, and demand an outlet for their own products. The result has been a rash of antitrust suits filed by companies which allege that vertical integration forecloses market entry and extends monopoly from telephone service to equipment manufacturing. Plaintiff suits against AT&T and General Telephone seek corporate divestiture between utility and supplier as relief to what they allege as market foreclosure (*7*).

Antitrust exposes the structure and conduct of the industry to public reassessment. It places the question of vertical integration on the policy front line. Telephone companies, of course, defend the structure, conduct, and performance of their industry with vigor and persistence. Nonetheless, the organizational philosophy, the way the industry has been put together, is experiencing unprecedented criticism. Accessibility to equipment markets thus represents a third area of conflict within the industry.

Causes of Conflict

What is behind this series of events described above? Why, after all these years, does the telephone industry find itself on the defensive? Why, after decades does an industry subject to public regulation find itself exposed to market competition? In a search for some answers, it is essential to reexamine the implicit assumption of the nation's telephone industry.

Telephone service, its equipment and organization rested on a presumption of technological exclusivity. That exclusivity originated with the patent right in 1876. It was subsequently extended by R & D and augmented by patent accumulation. It was later reinforced by policies that forestalled entry into equipment or telephone services.

Given technological exclusivity, it followed that market rivalry was deemed inappropriate. Rather, it was argued, the industry was endowed with what economists term "natural monopoly" characteristics. This meant the industry experienced declining unit costs as scale or size of plant increased (2, p. 973). "Natural monopoly" also suggested that the industry exhibited high capital spending that automatically precluded easy access or easy entry. Finally, telephone service was endowed with dimensions of public convenience and necessity. It followed, then, that exclusive supply, rather than market competition, would yield the benefits of efficiency, innovation, quality, and responsiveness.

Private monopoly inevitably issued the call for some form of accountability. Regulation evolved first at the state level and subsequently at the federal level (1, p. 5). Regulation as an institution was erected to assure that monopoly power translated into reasonable rates and adequate service. Regulation accepted the premise that competition was wasteful and inefficient. At the same time, regulation ensured that the monopoly firm should generate sufficient revenue to enable the firm to compete in the capital market. Technological exclusivity, in short, erected two institutions: the institution of private monopoly and the institution of public regulation.

It is that premise and those institutions that are now subject to question. The telephone industry, in the middle of technological explosion, finds that its expertise has migrated with a vengeance. Technological expertise in switching, transmission, equipment, manufacturing, and research and development no longer reside as the exclusive province of incumbent telephone carriers. The telephone industry is but part of the broad revolution besetting the electronics industry, a revolution that has spawned market competition.

Competition is a hallowed institution in the American scene. Antitrust laws governing corporate structure and conduct are long established. In the past, competition consisted of intraindustry firms competing for the consumers' dollar. Product substitutes added an element of competition to the market process.

But telecommunications technology is provoking a new economic phenomenon in the marketplace. Today, one can detect five industries, once separate and distinct, now moving on a collision course. These include the telephone industry, computer industry, semiconductor industry, terminal industry, and aerospace industry. This economic dimension goes beyond the competition of product substitutes, beyond competition within an industry. The telecommunications industry is experiencing *multiple interindustry competition*. This is a new species of market rivalry that is clearly unsettling to firms in a regulated environment.

The content of this multidimensional rivalry is the complexity and interaction of these five industries; the merging of markets represents a marked divergence from our past experience. Semiconductor firms compete with computer firms, telephone manufacturing firms compete with aerospace firms, and aerospace firms compete with data terminals. The permutations and combinations of this intense rivalry are indeed "mind boggling."

As a consequence of this technological turbulence, the technological diversity is seen to challenge the concept of universal telephone service, to challenge telephone manufactured equipment and to challenge holding company determination of R & D and investment. Technological migration is a fundamental challenge to the premise of scale economies, to the concept of wasteful duplication, to the notion of lower per unit costs, and to the very presumption of so-called natural monopoly.

Some assert, in fact, that technology challenges the very rationale of public regulation itself. Little wonder that the industry finds itself on the defensive. Little wonder that public policy now resides in the eye of a hurricane. Clearly two institutions, private telephone monopoly and public telephone regulation, are subject to change, reassessment, and reevaluation.

The strategic role of technological change is emphasized by the following illustration. Telephone companies find themselves attracted to the growing market of data communications. The question arises as to how carriers should diversify into these markets, whether they should file tariffs as regulated entities, or whether they should establish subsidiaries and enter the market as unregulated affiliates. Since 1971, the Federal Communications Commission (FCC) has ruled that non-Bell common carriers must establish separate affiliates for the purpose of selling data processing services on the premise that such services are competitive in nature and can sustain vigorous market rivalry (8).

In contrast, AT&T has filed tariffs on a data terminal that incorporates memory components (9). More questions arise: does memory constitute data processing? Is data processing appropriate for regulation? Do tariffs set a precedent for the extension of monopoly into markets previously regarded as competitive? That the FCC has presumably accepted the Bell tariff will undoubtedly precipitate a policy controversy over what are data communications, what constitutes data processing, what is the demarcation line between each, and whether such a line can be determined.

Finally, competition is intensifying in the domestic satellite market. A consortium under the aegis of IBM has received FCC approval to offer a digital satellite service within the United States. The system incorporating new microwave, satellite, and antenna innovations will undoubtedly compete with private line service offerings of common carriers.

But IBM's entry into data communications via a holding company raises a number of complex issues. Will such a consortium give IBM competitive advantages in software, hardware, protocols, data processing, and equipment compatibility not available to its competitors? And will IBM's entry into data communications eventually envelop the corporation with federal regulation as suggested by AT&T (10)? The fact that the Federal Trade Commission and the Department of Justice have called for further examination of these issues suggests that public policy remains unsettled and unresolved.

Public Policy Debate

The public policy debate focuses essentially on the service, equipment, and organization of the telephone industry. The issue has been joined with proponents advocating competition versus proponents advocating monopoly in this industry.

Service. First, there is the monopoly argument. Those supporting the monopoly approach assert that telecommunications is best secured through the vehicle of the regulated firms. Specifically, this position argues that exchange telephone service, intracity facilities, and even communication processing service is best rendered through a franchised utility with end-to-end system responsibility (*11*). This position is backed by several contentions. (i) Telephone service requires coordination and planning and must be rendered accountable to someone. That responsibility has traditionally fallen to communication carriers exclusively. The record of carriers in telephone service quality is eminently reassuring. Telephone service or quality in the United States broaches few rivals. (ii) Telephone subscribers gain through a policy of nationwide price averaging. Under this policy high profit subscribers subsidize low profit subscribers, low cost subscribers contribute to high cost subscribers. Both gain through the application of an overall averaging concept according to natural monopoly proponents.

Proponents for monopoly assert that rivalry, entry, and competition contradict the concept of system coordination. Competitive services result in wasteful duplication, cream skimming, and plant inefficiency. Technological imperatives and economic realities dictate that a single firm—the telephone carrier—be accountable to achieve the goals of efficiency, responsiveness, innovation, and quality to the subscribing public (*12*).

Let us consider ancillary services, such as value-added services or communication processing services. According to the telephone industry, such services pose a legitimate extension of traditional telephone service. Communication technology and transmission in switching and other service offerings lead inexorably to digital technology. The industry is on the forefront of that technology and has contributed to the non-voice as well as the voice state of the art. To deny the telephone industry an opportunity to grow in these new services would be arbitrary and capricious, and would flaunt the imperatives of technological change. Given the fact that communication processing services are regulated, the consumer is guaranteed that services to him are rendered at equitable and nondiscriminatory rates.

Second, there is the competitive argument. Proponents of competition assert that the "natural monopoly" argument is confined essentially to exchange telephone service. Here duplication or rivalry in central office exchanges can be wasteful, and no one advocates a policy of market entry. But natural monopoly has its limits; it is no longer applicable to intercity facilities and obviously is misplaced for data processing services.

In these new services, advocates of competition assert that market rivalry creates incentives that spur new services, new investment, lower costs, lower prices, and a broader array of consumer choice (*13*). Proponents argue that the marketplace best achieves the goals of efficiency and innovation, and that unfettered rivalry encourages the flow of resources into an industry and provides an impetus for the development of new markets.

Some proponents of competition insist that regulated carriers should be precluded altogether from markets congenial to entry, diversity, and rivalry. Movement by carriers into unregulated markets, they argue, carries with it an illegitimate extension of "natural monopoly." More important, diversification carries with it the illegitimate extension of government regulation. Once that regulation grows beyond the bounds of economies of scale, the pivotal barrier to entry thus becomes the very regulatory process itself.

Equipment. With respect to monopoly, telephone carriers consider themselves as the proper entities charged with equipment responsibility. Customer ownership of terminal devices can compromise a highly technical network, erode investment, dampen technological innovation, introduce cost inefficiencies, and degrade telephone quality (*1*). Only the carrier bears the continuing burden of equipment maintenance and repair. In the past, consumer benefits have resided in the concept of end-to-end service, and the telephone network must be protected from harm or "pollution" or it will be rendered useless to over a hundred million subscribers.

As to competition, its proponents assert that telephone technology has experienced a dramatic change. New equipment, new apparatus, and new hardware are spawned in a competitive marketplace. That process is aided and abetted by the incentives of technological rivalry. Competition and entry, in terminal apparatus, will alleviate capital shortage of equipment, reduce the construction burdens of the carriers, spur new equipment features, prompt lower prices, and enhance product innovation (*14*). Proponents of competition argue that standard interface requirements can protect the quality of the telephone network, that terminal apparatus exhibits none of the characteristics of "natural monopoly," and that the proliferation of data and computer terminals best proceed unfettered by regulation and unrestrained by restrictive tariffs.

Organization. Proponents of monopoly or advocates of holding company control of manufacture and supply believe that vertical integration must be judged on its performance. In their view performance has been outstanding. Integrated suppliers enjoy economies of scale, impressive productivity, low price, and innovation track records that redound ultimately to the subscribing public.

They say that equipment manufacturing must be coordinated and planned and be consistent with the objectives of overall system quality. Fragmentation of holding company control of buyer and seller increases cost, lowers quality, and compromises a nationwide telephone system. Holding company control of manufacture assures operating carriers that suppliers pursue profits not as an end in itself, but rather that integrated suppliers provide the right product at the right place and at the right time (*12*, pp. 9–29). Control of the manufacturing affiliates merely secures that end.

Proponents of competition argue that technological diversity has broadened R & D, broadened investment, broadened products, and broadened marketing expertise. Technology has also increased the number of firms capable of making contributions to the telephone equipment industry. Critics of vertical integration believe that the equipment market is not a natural monopoly, that it is not a regulated market, that suppliers are not public utilities. All manufacturers, they say, must meet the crucible of the market test in terms of equipment price, features, technology, and availability.

Proponents emphasizing the advantages of competition submit that market entry is the acid test of alleged economies of scale, that entry carries the virtue of economies of specialization and the economies of diversification. They argue that sole source manufacturing is a legacy of a bygone era, and that vertical integration compromises the nation's interest in R & D, in telephone equipment, in resource allocation, and in the incentives placed upon corporate efficiency and innovation.

Proponents insist finally that the equipment market has been foreclosed, and that the foreclosed market must be reopened so that the virtues and benefits of market forces can accrue to both the telephone operating companies and the telephone customers. In short, they imply that structural reform of vertical in-

tegration under holding company control is long overdue.

The reality of the policy debate. The debate on the pros and cons of service, equipment, and organization is no mere academic exercise. The telephone industry has found itself resisting and opposing service equipment ownership and the proposed changes in its organizational structure. In fact, the industry has introduced to the Congress legislation, titled the "Consumer Communications Reform Act" which marks a return to monopoly control of telephone service, telephone equipment, and of the telephone industry. This legislation, which would immunize the industry from competition, from antitrust, and from market rivalry, is controversial.

By contrast, competitors have challenged the carriers' position and contend that the carrier-sponsored legislation is merely an attempt to return an industry to operational methods of a bygone era. The public policy debate, monopoly versus competition, is thus very real. It is likely to gain intensity as technological migration continues to assault and question the premise of telephone carriers, its industry, and its regulation. Indeed, the debate and dialog are subjecting the very process of regulation to increased scrutiny and evaluation. The unprecedented rate of change in the telephone industry has resulted in a major debate as to the direction and content of public policy.

The Stakes of the Battle

It is clear from the turmoil in public policy that the stakes are enormous. To assess them requires an examination of certain trends in the economy and a determination of where these trends are driving us.

Consider first the trends in services. In recent years we have seen a proliferation of new services and new markets. Businesses can send mail electronically. Bills can be paid from the home over the telephone. Consumers can transfer money electronically, thus moving us toward a cashless economy. Important phone calls can be automatically forwarded. These new services are but the tip of the iceberg. Development and introduction of new services have become the hallmark of the industry. Marketing with emphasis on the specialized requirements of the customer has become a critical factor to corporate profits and organizational structure.

The second trend is the development of terminals. New services require the development of new terminals incorporating increased intelligence and memory. Already these terminals are becoming commonplace for such applications as remote banking, travel reservations, and point-of-sale devices for retail stores and supermarkets. Even the common typewriter is being replaced by word processing machines in business offices. By 1980 the annual number of terminals shipped is expected to increase from a present rate of 325,000 to 574,000 (*15*).

A third trend is the evolution of computer networks. New information services and remote intelligent terminals form the basis for growth in distributed computer networks. Thus, computing power is becoming more commonplace and is particularly important to such distribution oriented industries as warehouses and to branch facilities such as banks, retail stores, and supermarkets.

The trends in new services served by intelligent terminals and linked together with distributed computer networks are creating an exponential growth in data communications. For example, AT&T has estimated that data communications revenues will quadruple from $5.5 billion in 1975 to $22 billion in 1980 (*16*).

New services, new markets, and multi-industry competition are symptoms of a basic shift in the economic infrastructure of the United States. It represents a shift to an information economy. That we are in transition to an information economy there is no doubt; the evidence is abundant. Already some 50 percent of the U.S. work force is employed by the information sector of the economy, and U.S. service exports now exceed the exports of goods (*17*).

Just as the shift from agriculture to industry was in the last century, the shift to an information economy will be pervasive. The impact of this shift can be seen in how we produce, what we produce, how we transact business, how we pay, how we shop, how news is gathered and spread, where we work, what kind of work we do, and how we communicate. Information, its formation and transmittal, will play a key role in this new era. What we are witnessing is an information explosion covering the entire spectrum of our economy. It is not restricted to the computer or the communications industry. At the very least, it embraces banking, insurance, transportation, health, education communications, entertaining, and manufacturing.

Our dependence on information and its transmittal will be critical to the growth and prosperity of our economy. Policy decisions as to what information services are provided, who provides them and at what prices go to the heart of tomorrow's economic infrastructure. The stakes, markets, and investment are incalculable. It is in this context that the public policy issues surrounding telecommunications should be considered. And it is in this context that the merits of a policy of competition and monopoly in telecommunications should be debated and resolved. At issue is not merely the fortune of a single industry, but the thrust, content, and direction of our economy (*18*).

References and Notes

1. J. A. deButts, statement on behalf of AT&T before the U.S. House of Representatives Committee on Interstate and Foreign Commerce Subcommittee on Communications, 28 September 1976.
2. R. Nathan, *Hearings before the Subcommittee on Antitrust and Monopoly of the Committee on the Judiciary on S.1167* (Government Printing Office, Washington, D.C., 1974), part 6, p. 1974.
3. H. W. Boda, *Synergy: Technical Integration and Technological Innovation in the Bell System* (Bell Laboratories, Murray Hill, N.J., 1971), p. 46.
4. *1975 Annual Report of American Telephone and Telegraph Co.*, AT&T, New York, 1975.
5. U.S. Bureau of the Census, *Statistical Abstract of the United States 1975* (Government Printing Office, Washington, D.C., ed. 96, 1975).
6. "Coming: Another revolution in the use of computers," *U.S. News & World Report* (19 July 1976), p. 27.
7. International Telephone and Telegraph Corporation vs. General Telephone and Electronics and Hawaiian Telephone Company, Civil Action No. 2754 in the U.S. District Court, Hawaii (October 1967); United States vs. American Telephone and Telegraph Company, Western Electric, and Bell Telephone Laboratories, Civil Action No. 74-1698 in the U.S. District Court, District of Columbia (20 November 1974).
8. *Federal Communication Commission Reports, 28 FCC 2nd* (Government Printing Office, Washington, D.C., 1971), p. 267.
9. Datapoed 60/4 proposed tariff revision in American Telephone and Telegraph transmittal to the Federal Communications Commission, 12449 (1975).
10. In the matter of the application of Satellite Business Systems for authority to construct space segment and earth station facilities for use with its operational system at the 12/14 GHz frequency bands, and so forth, see file 7-DSS-P-76 for the reply of the American Telephone and Telegraph Company, 24 September 1976.
11. P. H. Henson, Testimony before the U.S. House of Representatives Committee on Interstate and Foreign Commerce, Subcommittee on Communications, 28 September 1976, p. 6.
12. J. R. Billingsley, "Values of vertical integration for the Bell system," *Sixth Annual Seminar on Economics of Public Utilities* (8 March 1973), p. 9.
13. P. M. Walker, Testimony before the U.S. House of Representatives Committee on Interstate and Foreign Commerce, Subcommittee on Communications, 29 September 1976, p. II-2.
14. C. G. Grant, testimony before the U.S. House of Representatives Committee on Interstate and Foreign Commerce, Subcommittee on Communications, 28 September 1976, pp. 25.
15. R. Salzman, *Datamation* (November 1975), p. 49.
16. "AT&T predicts data revenue increase," *Electronic News* (18 October 1976), p. 30.
17. Communication Engineering Digest, *Telecommunications*, May 1976, p. 22. The Conference Board, *The U.S. Economy in 1990*, February 1972, p. 66.
18. For additional references, see M. Irwin, in *Structure of the American Industry*, W. Adams, Ed. (Macmillan, New York, ed. 5, 1977); J. Brooks, *Telephone, the First One-Hundred Years* (Harper & Row, New York, 1975); Federal Communications Commission, "Investigation of the Telephone Industry in the United States" (Washington, D.C., 1939); R. F. Stone, M. A. Schusterman, C. G. Fenton, *Selective Competition in the Telephone Company* (T & E, Cambridge, Massachusetts, 1976).

International Trade in Electronics: U.S.–Japan Competition

John Walsh

Since World War II the United States has been the acknowledged world leader in electronics development. Despite stiff competition from Japan and Western Europe and shifts in the manufacturing of labor-intensive electronics products to less developed countries, the United States has maintained a favorable overall balance of trade in electronics. In the category of consumer electronics, however, the United States in recent years has experienced a substantial trade deficit and last year imports exceeded exports by some $3.3 billion (Table 1). Some observers also foresee erosion of the U.S. position in high-technology electronics exports, notably computers, an area in which American industry has been dominant.

Department of Commerce estimated figures for 1976 show that the trade balance for electronics declined from $2.2 billion in 1975 to $916 million last year.* The surge in the import of consumer products—particularly color television sets and citizens band radios (CB's)—accounted mainly for the reduction in the balance.

The major challenge to the electronics industry has come from the Japanese (*1*), and the most conspicuous penetration of the American market by Japanese manufacturers in the 1970's has been with color television sets. In an 18-month period beginning in 1975 the percentage of the American market held by Japanese manufacturers of color television receivers rose from 15 percent to about 40 percent. This had led to a series of actions by U.S. manufacturers and unions seeking relief under trade laws against Japanese exports. These actions are regarded as having serious implications because they come at a time when economic relations between Japan and the United States are at a particularly sensitive stage. Trends in electronics trade have also caused alarm because they appear to raise questions about U.S. competitiveness in world trade.

The size and diversity of the U.S. electronics industry—the Department of Commerce figures indicate the major categories—make it difficult to generalize about the industry as a whole. The statistical picture is complicated also by ownership of overseas manufacturing facilities by American companies and by the licensing and sale of U.S. technology, the proceeds from which are not clearly reflected in the Department of Commerce figures.

Nevertheless, it is possible to trace an evolutionary pattern in international trade in electronics. U.S. industry came out of World War II with the momentum imparted by wartime development of radar and sophisticated weapons and communications equipment. Continued heavy government spending on military and space electronics, with special emphasis on miniaturization and reliability, encouraged very rapid technological innovation. With a huge domestic market, the U.S. electronics industry has been geared primarily to producing for domestic consumption.

The Japanese have excelled in the quick application of electronics technology to consumer products. Japanese industry made early inroads with electronics exports because of low labor costs but demonstrated rapidly increasing abilities in high-quality engineering and the development of advanced manufacturing techniques and successful marketing strategies for U.S. and Western Europe markets. Japanese industry is heavily oriented toward exporting.

The "export psychology" of the Japanese is attributed essentially to the fact that Japan is a densely populated island chain faced with the need to import food, raw materials, and fuel. The Japanese response to defeat in World War II was to undertake the rebuilding of the country's industry and economic infrastructure along lines which would reestablish Japan as a strong competitor in world markets. The goal was achieved through an alliance of government, industry, and the work force which has inspired Westerners to use the term "Japan Incorporated."

A pivotal role in this alliance is played by the government's Ministry of International Trade and Industry (MITI). This agency does much more than promote trade in the ways that Western governmental bodies do; MITI sets priorities, gives direction, and provides various forms of assistance to particular industries and individual companies.

Most Western observers feel that special characteristics of Japanese economic structure and attitudes give Japanese firms an advantage in competing in foreign markets. Japanese firms, by and large, put less emphasis on profits than U.S. companies and are more concerned with growth and with obtaining a larger share of export markets. Banks in Japan often have special ties to particular companies which makes it easier for firms to obtain investment capital on more favorable terms than it is for Western firms. Employment practices in Japan, which stress "lifetime" employment, give Japanese companies the advantages of a stable, skilled, and motivated work force. Japanese law, including tax law, generally provides incentives for export activities. And government policy, historically, has been to resist upward adjustment of the value of the yen, which would tend to make Japanese products less attractive in world markets.

To an extent unmatched in socialist and Western countries, and certainly in the United States, Japan has been successful in planning and implementing foreign economic policy. By the middle 1960's the country had overcome balance-of-payments problems and after 1965 became accustomed to having its international balance of payments in surplus and to accumulating reserves of foreign exchange. During this period Japan began to transfer low-wage manufacturing activities in electronics and other industries to less developed areas such as Hong Kong and Singapore in the same way that American firms were.

In the early 1970's the Japanese

*In general, the United States is strongest in the export of computers and semiconductor devices and also in avionics and sophisticated military and communications equipment. The total value of electronic products in terms of factory sales in the United States in 1975 was put at $36.4 billion by the Electronic Industries Association (EIA). Although comprehensive figures are elusive, total employment in the electronics industry at large in 1975 was put above 1.1 million by the EIA. Current Department of Commerce estimates put the largest contingent of electronics workers in components manufacturing—some 355,000. The next largest group, about 316,000 workers, makes electronic systems for military and industrial uses. The computer and calculator category has some 200,000 workers. About 76,000 workers make consumer products.

The author is a member of the News and Comment staff of *Science*.

reached a kind of apogee of their international industrial success. But that success brought new difficulties. Industrial expansion had caused serious pollution and congestion problems. There was friction with the United States over textile imports and with European Community countries over electronics. And nations in Southeast Asia found Japan's trade and investment policies threatening.

In response to these pressures, the Japanese set out to formulate policies which would enable them to deal comprehensively with the problems of success. A notable effort was a draft report, "International trade and industry policies for the 1970's" (*2*), from the industrial structure council of MITI. The main recommendation was that national priority be given to knowledge-intensive and energy-economizing industry. The report urged that emphasis be placed on products of an R & D-intensive, high-value-added sort, such as computers, aircraft, nuclear-related equipment, specialized metals and ceramics, and on equipment which requires highly sophisticated manufacturing and assembly techniques, such as telecommunications gear, business machines, and pollution-monitoring instruments.

Such industry would be desirable in that it would not require heavy imports of raw materials. But the social dimension of the strategy is as important as the economic. The priority industries would be relatively free from pollution and would provide favorable work conditions for labor. The report appears to have captured the imagination of the public, even more so after the energy crisis of 1973, which had heavy psychological as well as economic impact in Japan. However, the proposal fell short of being enshrined as official policy, perhaps because of the criticism from foreign customers and competitors who might well have felt threatened. But most observers feel that the policy was tacitly accepted and is being implemented.

It is not surprising that Japan's export operations have led to frictions with other governments, not only with the United States but with Western European countries and with the European Community as an entity. The EC recently imposed a 20 percent levy on imported Japanese ball bearings, after an antidumping complaint by European industry was upheld (*3*). Negotiations are in progress over imports of Japanese automobiles, and talks on how to limit a virtual Japanese monopoly in shipbuilding have been going on under the auspices of the Organization for Economic Cooperation and Development in Paris.

In the United States the focus currently is on three actions aimed at halting Japanese penetration of the consumer electronics market. Most imminent is a decision on a petition by a group of unions and manufacturers claiming that the U.S. electronics industry has been injured by imports and requesting that quotas be imposed under provisions of the Trade Act of 1974.

The U.S. International Trade Commission, which is hearing the case, is a federal agency that does not act in a regulatory capacity but, rather, advises the President on what action, if any, should be taken. The President may or may not follow the recommendations.

A second case before the ITC is based on a petition by GTE Sylvania Inc., claiming that certain Japanese electronics companies are conspiring to dominate the U.S. market. The ITC has suspended its investigation of that complaint until it reaches a decision on the other case.

Of the actions now pending, the most far-reaching implications are attributed to a suit before the Customs Court in New York filed by the Zenith Radio Corporation against the Japanese on antitrust grounds. The focus of the complaint is a 15 percent commodity tax which Japan levies on domestically manufactured goods and then rebates on products which are exported. Zenith claims that the rebate amounts to a subsidy and violates a 19th-century U.S. law forbidding such subsidies. The issue is complicated because later international trade agreements seem to permit such rebates, since the payment of the tax falls on buyers rather than on manufacturers. If Zenith succeeds in its suit, a lot of other dominoes would presumably fall, as the value-added tax levied by the EC countries and indirect taxes imposed by Canada and other countries are similar to the Japanese impost.

The issue raised by Zenith is troublesome enough, but it is only one point of friction over trade between Japan and the United States. A decade has elapsed since major comprehensive trade nego-

Table 1. High-technology, electronics-oriented industries. U.S. exports and imports, 1974, 1975, and 1976 (millions of dollars).

Category	Imports				Exports				Trade balance			
	1974	1975	1976*	Change (%) 1975–76	1974	1975	1976*	Change (%) 1975–76	1974	1975	1976*	Change (%) 1975–76
Computers and related equipment	115	129	235	+82.2	2196	2224	2588	+16.4	+2081	+2095	+2353	+12.3
Calculating and accounting machines	361	336	335	−0.3	338	261	217	−16.9	−23	−75	−118	+57.9
Consumer products	2337	2059	3788	+83.9	384	392	498	+27.04	−1953	−1667	−3290	−97.4
Phonograph records and so forth	16	18	26	+44.4	57	59	66	+11.9	+41	41	+40	−2.4
Telephone and telegraph equipment	162	93	100	+7.5	160	198	227	+14.6	−2	+105	+127	+20.9
Commercial, military, and industrial equipment	505	456	653	+43.2	811	1012	1278	+26.3	+306	+556	+625	+12.4
Components	1304	1160	1644	+41.7	2267	1987	2532	+27.4	+963	+827	+888	+7.4
X-ray apparatus and so forth	121	173	199	+15.1	152	196	223	+13.8	+31	+23	+24	+4.3
Electronic measuring and test instruments	150	153	180	+18.4	438	495	492	−0.6	+288	+343	+312	−9.1
Totals	5071	4576	7160	+56	6803	6824	8121	19	+1732	+2248	+961	−57.3

*Figures for 1976 are estimated. [Source: Bureau of Census, U.S. Department of Commerce]

tiations were concluded among Western trading nations and Japan. Another cycle of hard bargaining is in prospect and, in the altered circumstances of the 1970's, some observers see the threat of a trade war.

From the standpoint of the U.S. consumer electronics industry, the major issue is the import of color television receivers. Framers of the petition to the ITC (*4*) say that during the first 7 months of 1976 imports of color sets rose to 1.3 million, an increase of 153 percent over the same period in 1975.

Recent years have also seen the establishment of a different sort of beachhead in the industry. In 1974 Motorola was sold to Matsushita Electric Industrial Company. And Sanyo has acquired a controlling interest in Whirlpool's Warwick division, which produces private brand sets for mass merchandisers.

The petition to the ITC emphasizes a decline in employment in U.S. manufacturing. Between 1966 and 1970, average employment in all U.S. television receiver plants is said to have dropped from 62,473 to 42,703, a decrease of 32 percent, while hours worked declined by 42 percent in the same period. The petitioners argue that the decline in sales of sets manufactured in the United States is attributable to the pricing of Japanese sets 10 to 15 percent lower than American sets and not to the superior quality of the Japanese sets.

Opponents in the action include Korean and Taiwanese interests as well as the Japanese—Matsushita, Sanyo, and Toshiba were the major Japanese manufacturers represented—and also the American Retail Federation (ARF). Major arguments by the opponents include the claim that a major factor in the drop in U.S. sales in 1974–75 was the recession and that automation in U.S. plants, not imports, were the main cause of the decline in employment. The lower cost of Japanese sets was attributed to advanced techniques in manufacturing and to an earlier move by Japanese makers to solid-state technology, which resulted in lower costs. AFR took the line that U.S. manufacturers had virtually ignored certain sectors of the market where there was potential consumer demand—particularly for small monochrome sets—and when U.S. manufacturers did offer such sets, they were of lower quality and more costly to maintain.

It is significant that the Electronic Industries Association, the industry's major trade association, did not testify at the hearings. A consensus view could not be developed within the membership because of the diversity of interests and viewpoints. For example, Zenith, the plaintiff in the Customs Court suit, is said not to favor quotas, a conventional protectionist approach to the problem.

To cast the industry position in straight protectionist versus free-trade terms is to risk oversimplification. U.S. policy has generally been to support liberalization of international trade—that is, liberalizing admission of foreign-produced goods to U.S. markets and vice versa—but subject to a large number of exceptions and hedges. U.S. textiles are protected by tariffs and quota agreements, for example. The domestic watch industry has traditionally been sheltered. And, although the United States is the leading exporter of agricultural products in the world, domestic wines and cheeses are protected.

Military Electronics

The export of military electronics has been a steady contributor to the favorable U.S. trade balance in electronics, but such exports raise problems of protection in a different context. Sales of sophisticated weapons and military aircraft are subject to rigorous controls, but the export of some other high-technology products and components and the licensing and sale of technology with possible military applications raises much more complex problems.

The most widely discussed recent statement on the subject was made in a report to the Defense Science Board on the export of U.S. technology by a task force headed by J. Fred Bucy, president of Texas Instruments (*5*).

The report concentrated on "know-how," rather than end products, except for items of prime military importance, and urged control of the transfer of technology by high-technology firms to non-Western and socialist countries and also cautioned against the retransfer of technology through U.S. allies.

Focusing on aircraft, jet engines, instrumentation, and solid-state devices, the task force argued that design and manufacturing know-how is essential to the maintenance of U.S. technological superiority. The report identified several levels of technology transfer, ranging from the most effective, "active" transfers, such as those involving turn-key factories with substantial training provided and the sale of licenses accompanied by "active teaching." At the other, "passive," end of the spectrum were the least effective forms of transfer, such as provision of blueprints unaccompanied by information that transmitted know-how.

The task force urged that this country design a policy to maximize lead time by preventing the transfer of strategic technology, particularly in fields where the United States has a significant lead.

In the consumer electronics field there seems to be fairly wide agreement that West-to-East transfer of technology has not been the crucial factor in Japanese success. The use by Japanese firms of U.S. manufactured components and the licensing of technology has doubtless accelerated progress. And some observers point out that some U.S. electronics manufacturers were allowed to establish manufacturing operations in Japan, where such initiatives by Western firms are usually discouraged. But Japanese performance is attributed to their choice of product, manufacturing, and marketing abilities. By and large, Japanese strategy is to choose a product for which technology is "mature," and for which a sizable consumer market exists, making volume production feasible.

From the middle 1950's on, the Japanese successively established strong, often dominant positions in the American market with radios, record players, monochrome television sets, tape recorders, and most recently color television sets.

Although the general pattern has been for Japanese products to displace U.S. products in the U.S. market, U.S. firms on occasion have used technological innovations to make successful commercial counterattacks. A major case in point was the competition for the calculator sales described in a 1975 Department of Commerce report (*6*). After World War II, U.S. firms employing electromechanical technology dominated the domestic market. By the middle 1960's, Japanese manufacturers exploiting electronics technology had become serious competitors. As late as 1966, U.S. manufacturers, still relying on old technology, held 90 percent of a $112-million market. By 1970, U.S. firms controlled 37 percent of a $224-million market.

In the later 1960's, Japanese calculator manufacturers began to use integrated circuitry largely supplied by American firms. By the early 1970's, several U.S. firms had concluded that logic and memory circuitry necessary for a four-function calculator could be included on a single integrated circuit chip. Rapid innovation made possible simplification of manufacturing, reduction in size, and cuts in labor costs. A headlong reduction in

price ensued, and when the cost of small calculators dropped below $100 a booming consumer market began to develop. American semiconductor manufacturers began to produce calculators rather than simply selling components (7), and as a result of this competition the Japanese share of the market dropped from about 40 percent in 1971 to 21 percent of an estimated $750-million market in 1974.

The drop in the average price of calculators in 5 years from about $400 to under $20 in 1974 was a dramatic example of price competition in the consumer electronics market powered by the rapid exploitation of technology. Some observers say that semiconductor manufacturers in Japan have caught up with American counterparts and that production of calculators is again shifting overseas.

The Japanese continue to invest resources in products which they believe have a substantial economic niche. New manufacturing capacity is created almost invariably only when expansion of exports is foreseen. The Japanese, for example, moved quickly and successfully to exploit the American vogue for CB radios, although there is no significant home market, present or prospective.

An even clearer case of export targeting is the development of the video tape recorder as a consumer product for home use. Under Japan's Electronic Industry and Specified Machinery Industry Promotion Temporary Measure Law (1971–79) the TV tape recorder was chosen to receive government direction of development, financial assistance, and special legal treatment, notably an exception from Japan's anticartel law. Video recorders were included with semiconductors and integrated circuits to receive assistance with cost reduction and commercialization. According to the Commerce Department study the Japanese are convinced that TV recorders will be a central factor in the next generation of consumer electronics. So substantial a lead have the Japanese taken that some Americans acquainted with the subject predict that no competitive TV recorders may be produced by American firms for the consumer market.

While Japanese dominance in the consumer electronics market has made American firms resentful—as the recent trade actions show—U.S. industry has tended to see this dominance balanced by the success of American firms at the high end of the technology spectrum, particularly in computers. But most observers now see the Japanese moving to challenge the American position here, too. Many are convinced that the Japanese are determined to establish an indigenous computer industry capable of competing across the whole range of products, including the biggest, so-called "mainframe" computers. There is general agreement that U.S. companies currently maintain a substantial technical lead in the computer field and hold the advantage in cost competition. But there is a widespread feeling that the long-existing status quo is about to undergo a drastic change. The source of the change is intervention in the field on the part of the Japanese government. The government is providing major support for development of semiconductors and computers. The aim appears to be for Japanese industry to supply domestic computer needs and to mount a challenge in international markets with a line of hardware, software, and services comparable to that offered by IBM and other major U.S. manufacturers. The scale of support is said to be about $200 million for the components field and a similar amount for computers. For some time, the government has backed a program to improve Japanese competence in software, where industry has been deficient. The government programs are aimed at industrywide strength and will utilize such devices as a common patent pool which is illegal in the United States.

Nascent Competition

As things stand, Japanese penetration of the U.S. market for computers and related equipment is not regarded as significant, although there was substantial rise in imports in that category last year. Japan, however, is regarded as having a foothold in the minicomputer field and is expected to compete in the market for the so-called desk-top computer, which is thought to have major growth potential. Japanese electronics firms have initiated relationships with American components and computer manufacturers which could lead to some technological leapfrogging, but the dominant impression in this country is that the Japanese intend to concentrate on internal development.

Some U.S. observers expect to see a repetition of events in the consumer electronics industry, with the Japanese initially scoring success with advances in semiconductors, probably with microprocessors. Others feel that U.S. firms hold the technological initiative and have shown that they can retain it in the semiconductor and computer fields. Interpretations tend to be highly subjective, since trends in this area are by no means yet clear. However, there is widespread agreement that a major factor in future developments will be the U.S. government's attitude toward the intervention of other governments on behalf of their own electronics industries. The obvious case is Japan, but the governments of Western European countries, notably France and West Germany, have taken major steps to foster domestic computer industries. These actions are based not only on the judgment that computers constitute a desirable growth industry, but that "intelligent machines" will increasingly affect virtually every aspect of modern life. Leaving security considerations aside, the governments of most major industrial countries take the view that to be without a viable, vertically integrated data-processing industry invites second-class status in the post-industrial period.

Up to now U.S. industry seems to have been tacitly willing to trade off imports in consumer electronics for exports in high-technology products. Now the U.S. government is being pushed by industry to reexamine its policy on the electronics trade in the face of big subsidies by other governments.

The questions raised are only beginning to be seriously discussed. In a sense, U.S. industry gained advantages similar to those provided by subsidies from heavy U.S. government spending on military and space electronics. Some analysts, incidentally, believe that the experience of certain American electronics firms in dealing with the government as their sole client has conditioned them to be poor competitors in international markets. And because U.S. firms have become habituated to expecting innovation to come only from within U.S. industry, they are said not to be adept at making use of the rapidly increasing number of important innovations being made abroad.

American firms are accustomed to dealing with the U.S. government in an adversary role as regulator and guarantor of competition. Competitive electronics firms here would like to see the U.S. government find a way to apply U.S. standards of fair competition to foreign firms which operate in U.S. markets. But dealing with the issue of the support by foreign governments of their high-technology industry which competes in U.S. markets is likely to be one of the thorniest issues confronting the Carter Administration in the next round of major trade negotiations.

References

1. Domestic and International Business Administration, Department of Commerce, *The United States Consumer Electronics Industry* (Government Printing Office, Washington, D.C., 1973).
2. *Proceedings of a Conference on Japan-U.S. Economic Policy* (American Enterprise Institute for Public Policy Research, Washington, D.C., 1975).
3. "Europe tries to hawk the Japanese," *Economist*, 12 February 1976, p. 99.
4. "Petition for import relief before the United States International Trade Commission," U.S. Customs Court, New York, 22 September 1976, p. 12.
5. "Highlights of the final report on export of U.S. technology" (presented to the Defense Science Board, Washington, D.C., 1976).
6. Domestic and International Business Administration, Department of Commerce, *The Impact of Electronics on the U.S. Calculator Industry 1965 to 1974* (Government Printing Office, Washington, D.C., 1975).
7. "Turning the tables on Japan," *Bus. Week*, 4 March 1972, p. 84.

Impact of Electronics on Employment: Productivity and Displacement Effects

Arthur L. Robinson

Structural unemployment is the technical name for the loss of jobs when changing technology makes currently held skills obsolete. If the question asked is "What impact will advances in electronics have on structural unemployment?", the answer is "Very little," according to most government economists and industry executives.

Nonetheless, individual workers will continue to feel substantial and sometimes traumatic impacts, including the need to learn new skills for jobs with different responsibilities and the necessity of relocating to other plants or to other parts of the country. Not only will the requirements for some old talents disappear and new ones be created, but, somewhat contrary to intuition, the electronics "revolution" may also open up opportunities for the relatively unskilled. In fact, the upcoming labor force may be structured into large numbers of relatively unskilled workers at one end and highly trained managers and engineers at the other, with very few medium-skilled individuals in between. And, as technological change continues at a rapid rate, a pattern of lifelong education and multiple careers will replace the tradition of training for one lifetime vocation.

According to some observers, the changes to be brought about by solid state electronics will, in the final analysis, be comparable to Henry Ford's move from custom to mass production. Since the invention of the transistor, the electronics industry has undergone a profound transition, and the changes of the last 25 years exceed those of the preceding 75. Moreover, there may be social impacts that even now are not fully appreciated. Some of these impacts most likely will involve the kinds of jobs that people work at and how many persons are needed for those jobs. If speculation about the electronics-dominated society of tomorrow turns out to be anywhere near the mark, then the impact of advances in electronics on employment could dwarf earlier concerns about automation, such as those typical of the early 1960's.

Hard numbers reflecting the impact of advances in electronics on employment are for the most part unavailable, in part because the problem is not yet of interest to manpower economists. Says one manpower specialist, "Displacement due to changing technology is now only a small part of the total job deficit, but the situation could change in the next two years if the recovery from the recession is slow and if technological change and competition from overseas increase." Technological change is also extremely difficult to separate from other influences, thus making its study unsatisfying.

One effect that can be documented is the large growth in the number of workers making and using electronics technology in the form of computers and communcations equipments. Marc Porat of the Department of Commerce's Office of Telecommunications has chronicled the rise of what some are calling the information economy (*1*). Included in the information economy are all jobs that involve the manufacturing of information machines, and the production, processing, transmission, distribution, or selling of knowledge or information. Also included are various information activities of both public and private bureaucracies, such as managing, planning, monitoring, marketing, and coordinating.

Not every occupation in the information economy depends, either directly or indirectly, on the existence of electronics technology, but a substantial fraction does. Porat's data show that, about 1955, a rising percentage of information jobs surpassed a declining percentage of industrial jobs, and that information activities became the dominant sector in the U.S. economy. From about 1965, the now dominant information sector accounted for about 45 percent of all employment (Fig. 1). Furthermore, as computers and information machines made their appearances in the late 1950's, exacerbating the trend toward large bureaucracies that began to build up in private industry and in government in the years after World War II, the percentage of the total national income earned by those in information occupations rose to more than 53 percent (Fig. 2).

Porat thinks that these trends offer considerable solace to those fearful of job displacements from electronics. In his view, every new computer or information machine introduced in an office, for example, leads to many more jobs than it displaces: in a more complex environment people are needed to gather information for the machine to process and to interpret what it puts out. A scenario of this type does, however, imply a steadily growing bureaucracy, and Porat worries how long society can support an increasing number of less and less productive individuals.

A more conventional viewpoint is that increased productivity arising from technological innovation will, given a reasonably healthy economy, create an increased demand for labor and raise real wages (*2*). C. Lester Hogan, vice-chairman of Fairchild Camera and Instrument Corporation, the third largest producer of semiconductor devices in 1976, says bluntly, "Advancing technology never reduces employment in the long run."

John Kendrick, chief economist of the Department of Commerce, pretty much echoes these sentiments. Kendrick, who has made a career of following the effect

The author is a member of the Research News staff of *Science*.

of technological change on labor productivity, recently told an IBM-sponsored seminar on this subject that "there is a significant positive correlation between industry rates of change in productivity and in output. . . . Thus, despite temporary problems of labor displacement, it would appear that technologically progressive industries create more jobs on net balance than the backward industries which are more vulnerable to dynamic change" (3).

Jerome Mark, assistant commissioner for productivity and technology of the Department of Labor's Bureau of Labor Statistics, explains why this might be so. Over all industries in the United States, there has been a relatively small spread in the hourly wage increases granted to labor, but there has been a much larger range in the growth of productivity. Thus, prices of goods are much more sensitive to productivity than to wages, and an edge in labor costs goes to the industry with the better growth in productivity. Reduced costs lead to more sales and ultimately more employment. Kendrick told *Science* that, far from causing displacements, technological change may even be having a smaller impact than it did 15 years ago because of a reduced research and development investment that has caused the rate of productivity increase to decline in the United States. Productivity advance is now lower in the United States than in some European countries and Japan, according to a recent article by economist Lester Thurow of the Massachusetts Institute of Technology (4).

The electronics industry has been on the front line of those feeling the effects of technological change, as well as being the source of the changes that affect the rest of the economy. It is perhaps also an exemplar of the innovation-increased productivity-lower costs-increased sales-more jobs scenario. During the period from 1960 to 1976, for example, about 664,000 jobs were added to electronics out of a total of 2,130,000 new jobs in all manufacturing industries, according to statistics provided by L. M. Rice, Jr., a vice president at Texas Instruments, the largest U.S. semiconductor device producer.

Most spectacular of all has been the rise of the U.S. semiconductor industry. From its beginnings in the early 1950's, the industry mushroomed to sales of about $520 million in 1961 (the era of the discrete transistor) and to 1976 sales of about $2.4 billion (the era of the integrated circuit). The worldwide industry (excluding the Communist bloc) sales were about $5.4 billion (see Fig. 3).

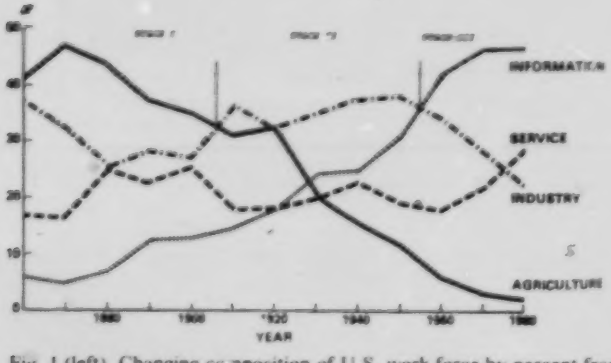
In 1961, between 10,000 and 15,000 employees produced about 520 million devices, nearly all of which were discrete transistors, diodes, and the like. By 1976, about 120,000 workers made a total of about 6 billion devices. (A device could be a discrete transistor or a complex integrated circuit containing 20,000 transistors; but, on the average, according to Fairchild's Hogan, who supplied the figures, a device contained about 50 transistors.) Thus, while the average labor productivity measured in devices per employee did not change greatly, measured in "transistors per employee, the average labor productivity skyrocketed. And the increased sales stimulated by the resulting lower cost was the foundation for the more than a tenfold increase in employment during those 15 years.

At Fairchild, experience with manufacturing silicon diodes provides another perspective of the effect of productivity. In 1966, Fairchild sold $87 million worth of these devices with about 200 employees to make them. By 1976, sales had decreased somewhat—to $50 million; but the number of employees had increased slightly to 225. The startling statistic is that the number of labor hours required to manufacture 1000 diodes had decreased by a factor of 10. Therefore, the increased productivity had stimulated sales to about 10 times the number of diodes sold previously. In this instance, however, increased productivity did not greatly increase the level of employment, but did perhaps postpone the demise of a device that is being supplanted by other types.

Not everyone is quite so euphoric. One who is not is J. A. Haddad, vice president for engineering programming and technology at IBM. Haddad feels that careful and detailed econometric studies are necessary before expansive projections become anything other than speculation. The IBM executive points to two examples, one of which vividly backs up the productivity and growth argument and one of which does not.

In the former category is the well-trod example of the hand-held calculator, which was made possible by the development of inexpensive but computationally powerful general purpose integrated circuits. While makers of slide rules may have taken a beating, relatively few persons ever bought slide rules in the first place. The hand-held calculator, on the other hand, has so caught the public fancy that practically "every kid has one now." The volume of calculators produced has probably more than made up for the number of displaced slide rule manufacturers and their employees.

Digital watches are a different story. Mechanical watches were already pervasive among all groups of society and probably not that many people own more than one or two. Since relatively few man hours are needed to make a first-



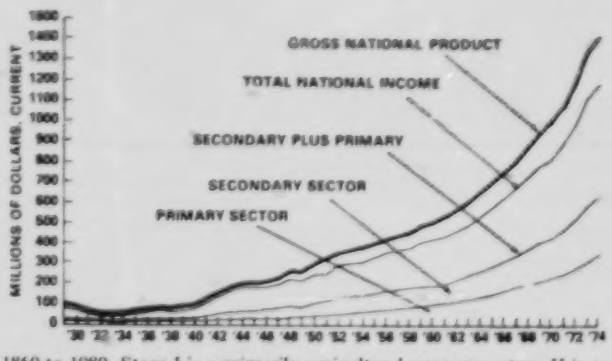


Fig. 1 (left). Changing composition of U.S. work force by percent from 1860 to 1980. Stage I is a primarily agricultural economy; stage II is an industrial economy; and stage III is the information economy. Fig. 2 (right). Portion of total national income originating in the information sector, 1929 to 1974. The information sector comprises a primary (manufacturing, information producing, information processing and transmission, and information selling) sector and a secondary (information services produced and consumed internally by industries and government) sector. [Source: M. U. Porat, Office of Telecommunications]

class digital watch. Haddad guesses that the net effect of electronics in this case may be to decrease employment.

A concern of many nowadays, including Haddad, involves the question of how much longer society can continue to rely on unlimited economic growth. "Sooner or later, continuing economic growth will no longer be sustainable because of dwindling natural resources and supplies of energy," Haddad insists. "We will be as big as we can get." At this point, the question will be one of finding how to improve the standard of living without further growth. Other industry executives would counter Haddad's argument by saying that they never meant that growth is the same as increasing consumption. Says William Hittinger, executive vice president for research and engineering at RCA, it is a matter of "growth by new concepts, not just consuming more."

Whatever the outcome of debates about economic growth, it is clear that within the electronics industry, changing technology has already effected major changes in employment of a qualitative nature in addition to the quantitative one already discussed. From a different viewpoint, RCA's Hittinger restated some of Porat's conclusions about technology leading to increased bureaucratic activities. According to the RCA research director, solid state technology of all types, not just silicon integrated circuit technology, has caused an expanding demand for people of many backgrounds and skills.

After World War II when vacuum tubes were still the dominant active devices, electronics companies mainly provided an off-the-shelf, catalog service for discrete devices that other companies used to build their products, and most employees had physical science backgrounds. Today there is a full spectrum of practitioners of basic sciences (physics, chemistry, metallurgy) and applied sciences (electrical and mechanical engineering, statistics), and there is, as well, a great demand for the "softer" sciences (sociology, economics).

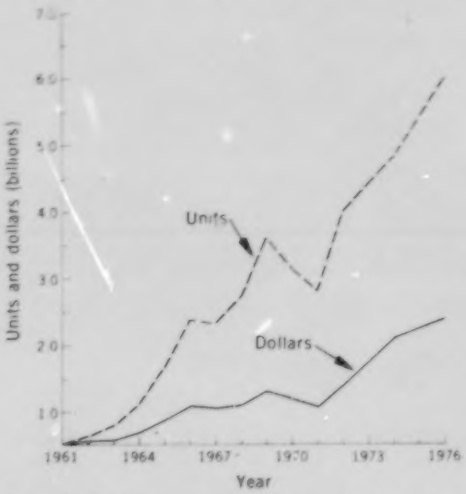
The need for the softer science practitioners has arisen because semiconductor houses are now marketing products and systems of their own, and must therefore engage in such market research activities as demographics, consumer preference testing, and input-output studies. If your company's microprocessor is the essential ingredient of a pocket calculator, why sell the microprocessor to someone else to package? Sell the complete product yourself.

There has also been change within the manufacturing divisions of semiconductor companies, as the transition from discrete devices to integrated circuits transpired. Hollis Caswell, director of applied research at IBM's Yorktown Heights laboratory, describes some of these changes. Fifteen years ago, an engineer designed a silicon transistor with a slide rule (chose the vertical dimensions, the specific impurities and their doping levels, and the geometry needed to give the desired characteristics); he had the transistor fabricated; and he tested it. He could do this and not be overly concerned with the interactions of the transistor with other discrete components of his circuit.

Now with silicon integrated circuit technology, building individual transistors and designing logic or memory circuits merge into the same task. Furthermore, the engineer will design the circuit with the help of a computer with which he can interact by way of a light pen and a cathode ray tube display. He will simulate the circuit performance on the computer, a process known as software bread-boarding, in contrast to the previous practice of wiring transistors together into a circuit to see how the assembly worked.

Similarly, the process of making the series of masks used in the integrated circuit fabrication process has been computerized. Where once the engineer would actually draw patterns on a piece of paper, the patterns being photographically reduced after being manually transferred to a suitable working material, today the engineer and his interactive computer can bypass much of the manual labor. In fact, according to Hogan at Fairchild, much of the work can be carried out by technicians trained to use the computer once the basic logic design for the circuit has been worked out.

As does IBM, the Western Electric Company, which is the manufacturing arm of the Bell System, itself makes a substantial fraction of the electronic devices needed for its products: telephones, switching machines for branch and for long distance (toll) offices, and other equipment. The pervasiveness of advanced electronics in the telephone system was the subject of a recent speech to Bell Laboratories officials by Joseph West, vice president and chief operating officer of Western Electric. The theme of West's presentation was that, although advancing technology is the lifeblood of the Bell System, it is not



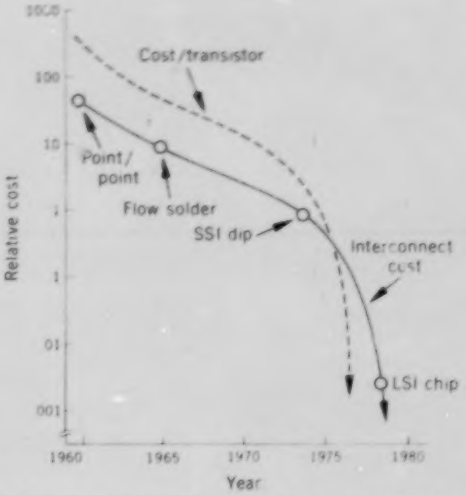


Fig. 3 (left). Semiconductor industry sales by year for the period 1961 through 1976. A unit is the semiconductor device defined in the text. Over the period graphed, the average price per unit sold declined from $1.00 to about $0.32. Since the average number of transistors per unit increased from 1 to about 50, the cost per transistor plummeted by a factor of 150. [Source: C. L. Hogan, Fairchild Camera and Instrument Corporation] Fig. 4 (right). Cost trends of transistors and interconnections at Western Electric. Point/point refers to connecting discrete transistors on a panel. Flow solder refers to connecting a transistor into a circuit on a printed circuit board. SSI dip refers to an estimated cost for the interconnection between transistors in an early small-scale integrated circuit. LSI chip refers to the same cost for a contemporary large-scale integrated circuit. [Source: J. T. West, Western Electric Company]

Table 1. Impact of changing technology upon the Switching Division of the Western Electric Company. [Source: Western Electric Company]

Year	Employees at year end		Switching lines shipped	
	Total corporation	Switching division	ESS	Total
1962	144,874	26,962	1,000	2,723,100
1963	141,468		5,000	2,734,100
1970	178,000	39,000	1,728,000	5,400,000
1975	147,651		3,386,000	4,358,200
1976	144,123	27,172	4,683,000	5,473,000

without certain problems for those managing its implementation.

Western Electric experience with the transition from discrete semiconductor devices to integrated circuits has paralleled that of the semiconductor industry. But because Western Electric is primarily concerned with the manufacture of end products, there have been other effects. The manufacturing of devices, for example, has taken on a substantial portion of the manufacturing effort that was formerly carried out in the final assembly operation. This has occurred because so much of the circuit interconnections are incorporated as an essential part of the integrated circuits now being manufactured. Thus, says West, "where historically the task of interconnecting circuit elements has been very labor intensive, as we move more and more into large-scale integration, we see an interconnect cost trend which closely tracks that of the cost per equivalent transistor" (Fig. 4).

Switching machines to route telephone calls through branch exchanges and long distance toll offices are a major product of Western Electric and indeed are central to the telephone business. They also provide some excellent illustrations of the impacts of technology. Shown in Table 1 are some statistics for production and employment in Western Electric's Switching Division over the last 15 years.

At the start of this period, the dominant switching technology was an electromechanical one generally known as crossbar. In 1962, the first all-electronic switches were manufactured under the appellation ESS. The most recent versions of ESS make extensive use of integrated circuitry and are heavily computerized—that is, their functions are determined by stored programs rather than by fixed or "hard-wired" circuits.

West told the Bell Laboratories officials, "An unglamorous way of describing your new designs in manufacturing terms is to point out that you have designed for us a switching machine that in terms of lines switched weighs one-sixth as much as [crossbar], has one-fifth the connections, takes one-third the floor space to make, and somewhat surprisingly, takes one-third the capital and substantially less installation and engineering. Direct impact can be seen on Western's need for wiremen—installers—standards engineers."

But the impact on Western Electric's need for switching division employees has been more qualitative than quantitative. Twenty-five years ago, according to Western Electric officials, a typical communications equipment factory comprised three categories of operations: (i) materials working using punch presses, molding presses, and automatic screw machines; (ii) apparatus assembly and adjusting; and (iii) equipment assembly, wiring, and testing. Cable forming and manufacture of vacuum tubes also were part of the factory.

Nonetheless, displacement effects due to the changeover to integrated circuit technology with more chemically based processes and more computer aided processes were not large. From 1962 to 1976, the number of workers there did not change greatly; indeed total Western Electric employment did not change much, but the total number of switching lines doubled. Furthermore, a projection further into the future reveals that the demand for switching lines may continue to rise but without a need for additional division employees.

As compared to switching division employment in 1970, the lower numbers of employees in 1976 would seem to indicate dramatic displacement effects. But for various reasons company officials regard employment that year as anomalously high. Also to be borne in mind is that not all division employees are directly involved in manufacturing switches.

That the qualitative nature of jobs will change with different skills, different levels of training, and different responsibilities being involved has other repercussions. In addition to a 44 percent decrease in employment in the Switching Division between 1970 and 1976, there was, according to Western Electric officials, a marked shift in the distribution of personnel over grade levels. Where once there had been a smooth distribution in hourly skills, gradually decreasing in numbers at higher skill levels, now the curve shows two pronounced peaks. The first occurs at high grade levels and reflects the need for highly trained testers and analysts. The second peak occurs at the lower end of the range of skills. A rather large valley corresponding to decreased requirements for middle level skills rests between the two peaks (Fig. 5, a and b).

At IBM, Haddad, who more or less seconds this perception, points out that the thrust of technological advance is generally to "take out the middle." Sophisticated levels of training are needed to design and less training is needed to "make things happen." Haddad philosophizes that this trend is part of an overall drift in society that is taking away from the concept of the whole, self-sufficient man. Everyone, whatever the skill and level of training, is becoming more interdependent.

Most of the individuals who will feel the impact of the advances in electronics, no doubt, do not work for the semiconductor industry nor for companies like IBM and Western Electric. Thus, the effects of electronics on employment in these industries may not be entirely representative; they just happen to be the easiest to document for the present. One example, possibly more representative, resides in the experience of the operating telephone companies, those which use the switching equipment manufactured by Western Electric. The data in Table 2 show that the introduction of the all electronic switching equipment in long distance toll offices is estimated to have reduced the need for administration and maintenance people to from less than one-half to less than one-third, depending on the size of the office. The reduction is due in part to the availability of computerized diagnostic and testing systems for use in administration and maintenance. Where once an administrator was needed to order maintenance work after detection of a malfunction, now detection and decision-making can be done electronically.

The mail (5), banking (6), offices (7), schools (8), factories (9), and communications industries (10) are some of the places where electronics will change employment patterns both qualitatively and quantitatively, although in most cases the impact is as yet more speculation than fact.

At the Bureau of Labor Statistics,

however, two studies were made in the early 1970's of the effects of computerization, one on electronic data processing machine makers and users (*11*) and one on process control in factories (*12*). According to Mark at the bureau, there was no dramatic evidence for widespread displacements in either area, although both studies were completed before minicomputers had become as pervasive as they are now. Mark did, however, support the observations made at Western Electric and IBM as to the changing distribution of skills needed. He added the further observation that the grade levels comprising the lower peak of the distribution approximately corresponded to semiskilled, not unskilled, personnel. In addition, there was some demand for mid-level skills in the form of repair and maintenance people.

Martin Ernst of Arthur D. Little, Inc., thinks that retail store automation systems will be among the first to have significant employment effects. Point-of-sale terminals in retail and grocery stores have the potential of enabling automatic checkout and inventory management with the consequent lessening need of labor for these functions. Connection of terminals in stores directly to computers in banks and other financial institutions and of computers in financial institutions to one another may also lead to a less cash-oriented society with less need for persons to process financial paperwork. Ernst guesses, however, that what personnel reductions do occur will be by attrition—failure to rehire—rather than by firing or layoffs.

The United States Postal Service is another likely candidate for large-scale impact by electronics. With about 650,000 employees, the Postal Service is the third largest nongovernment employer in the United States. Currently, it is in big trouble, as rising costs and decreasing revenues are leading to larger and larger deficits ($1.2 billion in fiscal year 1976) and the prospect of poorer and poorer service. (Surpluses in the second half of 1976 are regarded as unreliable indicators of the long-term situation.) Both the public and the U.S. Congress, which supports the mail organization with subsidies, are screaming for change. At present a congressionally established Postal Service Commission is feverishly putting together a set of recommendations on what should be done; its report is due by the end of the month.

One of the reasons for decreased revenues—a reason that most likely will become ever more important if postal service declines further—is that electronic technologies are permitting users to find alternative ways to send first class mail. Electronic transmission of messages, electronic transmission of funds, and data transmission systems will soon be able to service most of the needs of commerce; and business mail accounts for 74 percent of the nation's letter mail. A just completed National Research Council study on electronic message systems for the U.S. Postal Service estimated "as much as 30 percent of all government and business mail may be amenable to transmissions as electronic messages" (*13*).

Table 2. Estimated impact of changing technology on the operating telephone companies of the Bell System. [Source: Western Electric Company]

Type office	Administration and maintenance people required	
	ESS	Crossbar
Small toll (20,000 trunks)	32	68
Large toll (100,000 trunks)	60	200

Interestingly enough, in the Postal Service's 6-year life, this is the fifth report recommending the adoption of electronic message transmission by the Postal Service (a sixth is still under way), yet practically no planning in this direction has begun. In this case, even the labor unions are clamoring for the Postal Service to adopt electronics technology. J. Joseph Vacca, president-elect of the 230,000-member National Association of Letter Carriers testified to the Commission on Postal Service that "I am firmly convinced, though that if we do not soon become . . . comfortable with the language of Electronic Message, Fund and Data Transmission, we will be literally out of business" (*14*).

Vyra Imondi, director of electronics communications studies on the commission, points out that approximately half of the current Postal Service employees will be eligible for retirement in the next

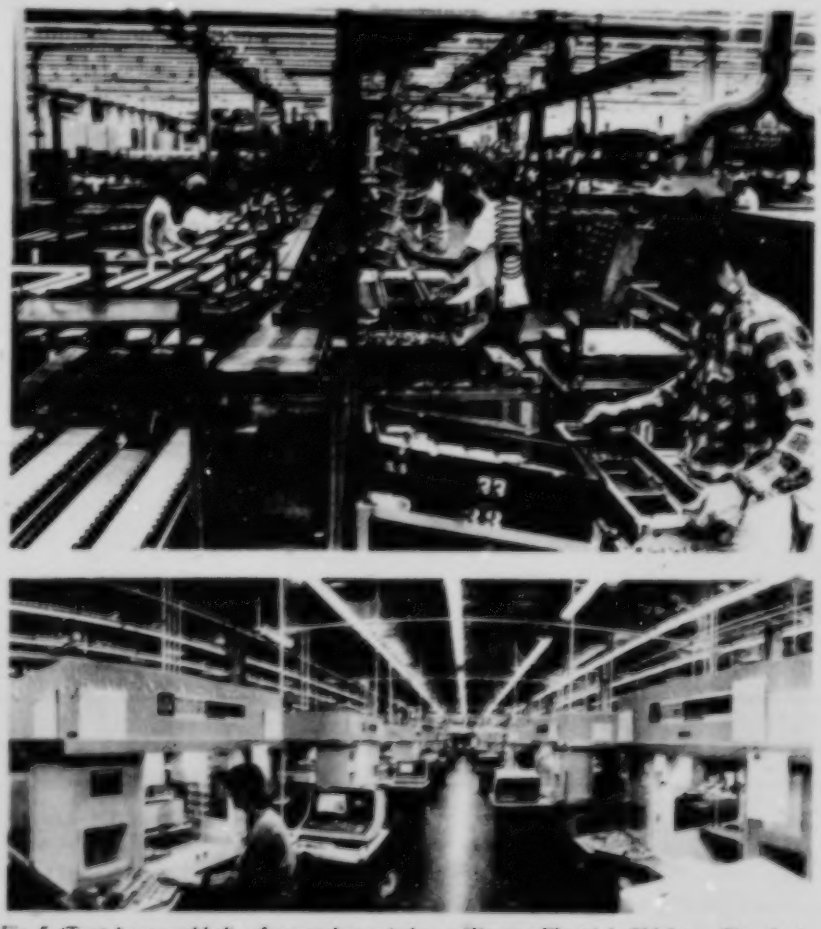


Fig. 5. (Top) An assembly line for crossbar switches at Western Electric's Oklahoma City plant. The assembly line typified the electromechanical crossbar technology of 15 years ago. (Bottom) Assembly area of Oklahoma City plant today. The operation illustrated is the wiring of grids on multilayer printed circuit boards. [Source: Western Electric Company]

10 years. If the Postal Service adopts the new technology, many will be hired with new skills to fill in for those retiring. But the failure to adopt electronic mail technology made possible by advancing electronics could be the driving force for the further deterioration of postal service and the concomitant loss of jobs.

From the point of view of the individual who is impacted by advancing electronics technology, the most immediate questions are personal. Will I lose my job? What provisions are there for me to retrain into another job in the company? Will I be transferred to another location? To the individual faced with these and other uncertainties, euphoric pronouncements about productivity, multiplier effects on employment, better standards of living, and the like, are pretty irrelevant.

Economists and some industry executives view the problem as one easily solved by retraining. Says the Office of Telecommunication's Porat, "Displacements are important to economists only if people are not retrainable and if phasing out of occupations is not anticipated. One wants labor resources to shift as soon as possible because it is inefficient for them to be idle." Furthermore, the very technology that may cause temporary displacements can be used to ameliorate their effects. For example, two-way computer aided instruction sent into the home or neighborhood learning centers by way of satellites or broadband cables could be one vehicle for retraining. Porat underscores the desirability of the now developing enthusiasm for lifelong learning as being especially important for times of low population growth, when lateral mobility will be the major kind.

Retraining has become a big business at Western Electric. In the next several years, the company will put into service some 800 electronic switching systems. About 20 percent of the company's installation efforts are now in this area, but this figure will increase as more telephone systems become all-electronic. In former years, Western Electric would generally not retrain installers experienced in electromechanical technology, but would start from scratch by training new personnel. Now, as crossbar demand slackens, the company has a near-crash program under way to retrain electromechanical people to proficiency in ESS installation. Since 1975, about 1000 such people have received some ESS instruction (*15*).

Widespread retraining is not always needed. Says a company official, "Displacement effects in the Switching Division were small because attrition was such in the crossbar operations that limited hiring was required to maintain a crossbar work force" in the 1960's. Retraining was necessary, however, to "develop electronics diagnostic capabilities in the testers. This was accomplished in a number of ways, including in-hours classroom training, after-hours company sponsored schools and on-the-job training."

Priding itself on not having to lay off employees, IBM also has extensive retraining programs. But, points out Caswell, "of those having skills needed 15 years ago but no longer needed, many are retrained into sales, sales support, and service jobs, not other manufacturing jobs." Haddad observes that retraining is very effective, but its usefulness varies with skill level. People with Ph.D.'s are generally flexible and can do many jobs from the start, so retraining needs are minimal. On the other end of skill curve, retraining is very effective because there is not that much skill to be learned. Even unskilled people off the street can be quickly trained for many electronics manufacturing jobs. It is the mid-level people who are hardest to change. A person with a lifetime devoted to mechanical skills just does not pick up electronics the way a younger person raised in the era of calculators does, according to the IBM executive.

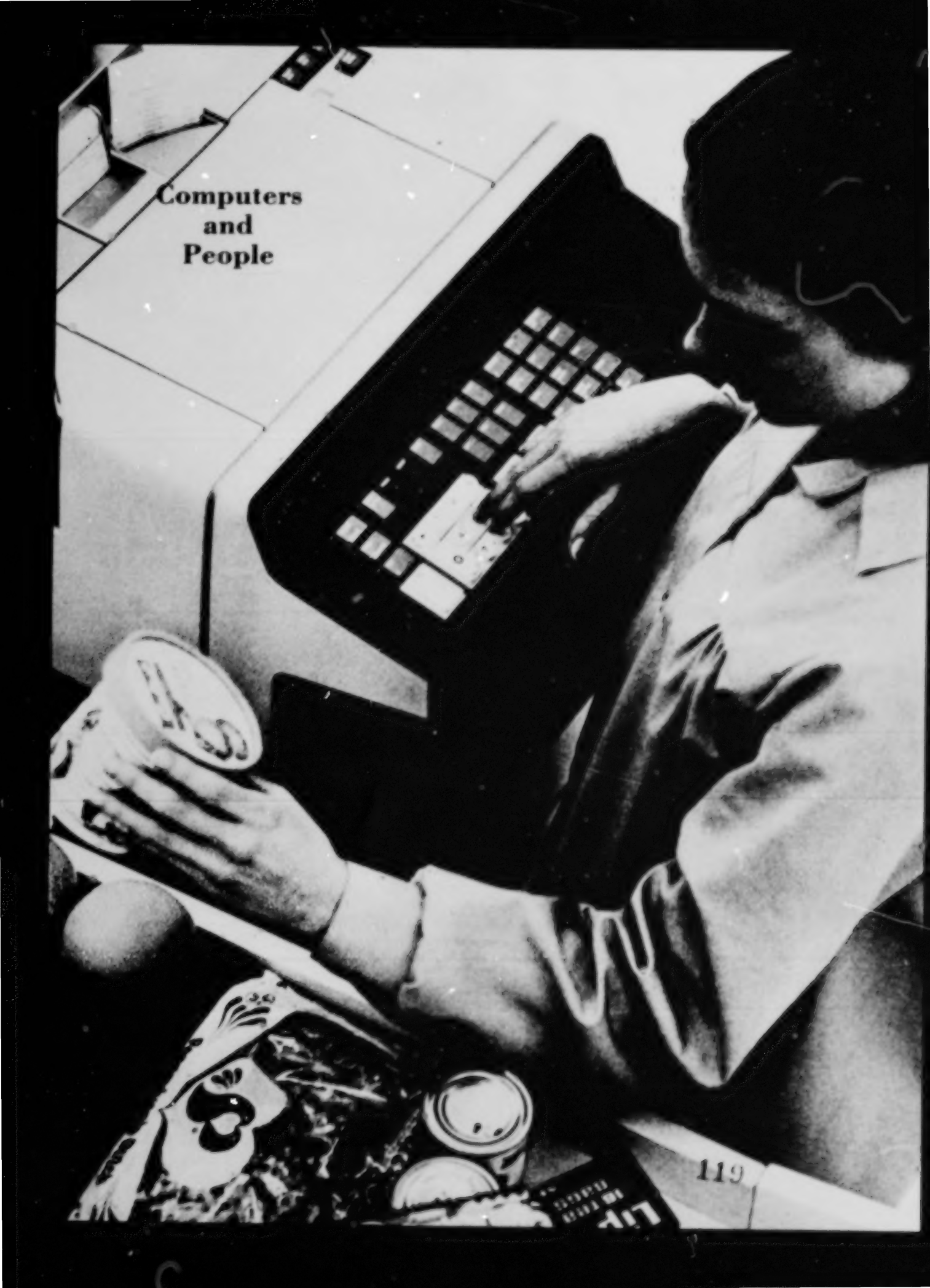
For precisely this reason, as well as for many others, labor union officials do not regard retraining as the panacea for all displacement due to technological change. According to Markely Roberts, an economist in the department of research of the AFL-CIO, there is "no single solution, but a whole variety of approaches. Which is best depends on the industry and the quality of the labor-management relationship." Roberts has summarized some of the possible strategies for dealing with displacement effects (*16*). In addition to retraining, he included prior notice, attrition, seniority rights, early retirement, transfer and relocation rights, and severance pay as possible mechanisms to cushion labor from the effects of change. In all cases, the labor philosophy is that the companies should regard the costs of cushioning labor as one of the expenses of introducing new technology, and these costs should be just as important in deciding upon the economic viability of new technology as any other. Roberts concluded, however, that the best protection for labor resided in a full employment economy.

Another tactic will be tried by the Communications Workers of America in its upcoming negotiations with the American Telephone and Telegraph Company (AT&T). Since 1974, AT&T has reduced its work force by almost 100,000 to a current total of about 900,000 employees. The communications workers want the Bell System to spread the available work by reducing each person's work hours without a cut in pay (*17*).

Few of those who ought to be concerned with the effect of electronics on employment seem to think of it as other than one more step in the evolution of technology. Such phenomena as the rise of bureaucracies, simultaneously made necessary and possible by communications technology, and the lumping of worker skills at the ends of what was once a smooth distribution would seem to indicate that more than evolution is going on. The evolution not revolution viewpoint may well be correct, but if it is wrong, a long line of unemployed may find themselves with the wrong skills for the wrong job.

References

1. M. U. Porat, *The Information Economy*, Vol I, thesis, Program in Information Technology and Telecommunications, Center for Interdisciplinary Research, Stanford University, 1976.
2. H. A. Simon, *Science* **195**, 1186 (1977).
3. N. Valéry, *New Sci.* **73** (1034), 70 (1976).
4. L. C. Thurow, *Newsweek*, 14 February 1977, p. 11.
5. R. J. Potter, *Science* **195**, 1160 (1977).
6. F. E. Balderston, J. M. Carman, A. C. Hoggatt, *Science* **195**, 1155 (1977).
7. M. Shepherd, Jr., speech at *Trends and Applications Symposium: Micro and Mini Systems*, National Bureau of Standards, Gaithersburg, Maryland, 27 May 1976.
8. J. F. Gibbons, *Science* **195**, 1139 (1977).
9. N. H. Cook, *Sci. Am.* **232** (2), 23 (1975); L. B. Evans, *Science* **195**, 1145 (1977).
10. D. Farber and P. Baran, *Science* **195**, 1166 (1977); M. R. Irwin and S. C. Johnson, *ibid.*, p. 1170.
11. U.S. Department of Labor, Bureau of Labor Statistics, *Computer Manpower Outlook*, Bulletin 1826 (Government Printing Office, Washington, D.C., 1974).
12. U.S. Department of Labor, Bureau of Labor Statistics, *Outlook for Computer Process Control*, Bulletin 1658 (Government Printing Office, Washington, D.C., 1970).
13. National Academy of Sciences, *Electronic Message Systems for the U.S. Postal Service* (National Technical Information Service, Springfield, Virginia, 1976).
14. Testimony given before the Commission on Postal Service, 28 January 1977, Washington, D.C.
15. T. Dau, *WE* **28** (6), 2 (1976).
16. M. Roberts, *American Federationist* **80**, 13 (1973).
17. "Union to Demand Job-Security Provisions in AT&T Pact to Spread Available Work," *Wall Street Journal*, 16 February 1977, p. 7.


Computers
and
People

What Computers Mean for Man and Society

Herbert A. Simon

Energy and information are two basic currencies of organic and social systems. A new technology that alters the terms on which one or the other of these is available to a system can work on it the most profound changes. At the core of the Industrial Revolution, which began nearly three centuries ago, lay the substitution of mechanical energy for the energy of man and animal. It was this revolution that changed a rural subsistence society into an urban affluent one and touched off a chain of technological innovations that transformed not only production but also transportation, communication, warfare, the size of human populations, and the natural environment.

It is easy, by hindsight, to see how inexorably these changes followed one another, how "natural" a consequence, for example, suburbia was of cheap, privately owned transportation. It is a different question whether foresight could have predicted these chains of events or have aided in averting some of their more undesirable outcomes. The problem is not that prophets were lacking—they have been in good supply at almost all times and places. Quite the contrary, almost everything that has happened, and its opposite, has been prophesied. The problem has always been to pick and choose among the embarrassing riches of alternative projected futures; and in this, human societies have not demonstrated any large foresight. Most often we have been constrained to anticipate events just a few years before their occurrence, or even while they are happening, and to try to deal with them, as best we can, as they are engulfing us.

We are now in the early stages of a revolution in processing information that shows every sign of being as fundamental as the earlier energy revolution. Perhaps we should call it the Third Information Revolution. (The first produced written language, and the second, the printed book.) This third revolution, which began more than a century ago, includes the computer but many other things as well. The technology of information comprises a vast range of processes for storing information, for copying it, for transmitting it from one place to another, for displaying it, and for transforming it.

Photography, the moving picture, and television gave us, in the course of a century, a whole new technology for storing and displaying pictorial information. Telegraphy, the telephone, the phonograph, and radio did the same for storing and transmitting auditory information. Among all of these techniques, however, the computer is unique in its capacity for manipulating and transforming information and hence in carrying out, automatically and without human intervention, functions that had previously been performable only by the human brain.

As with the energy revolution, the consequences of the information revolution spread out in many directions. First, there are the economic consequences that follow on any innovation that increases human productivity. As we shall see, these are perhaps the easiest effects of technological change to predict. Second, there are consequences for the nature of work and of leisure—for the quality of life. Third, the computer may have special consequences for privacy and individual liberty. Fourth, there are consequences for man's view of himself, for his picture of the universe and of his place and goals in it. In each of these directions, the immediate consequences are, of course, the most readily perceived. (It was not hard to foresee that Newcomen's and Watt's engines would change the economics of mining in deep pits.) It is far more difficult to predict what indirect chains of effects these initial impacts will set off, for example, the chain that reaches from the steam engine through the internal-combustion engine to the automobile and the suburb.

Prediction is easier if we do not try to forecast in detail the time path of events and the exact dates on which particular developments are going to occur, but to focus, instead, upon the steady state toward which the system is tending (*1*). Of course, we are not so much interested in what is going to happen in some vague and indefinite future as we are in what the next generation or two holds for us. Hence, a generation is the time span with which I shall be concerned.

My discussion will be divided into five parts, the last four corresponding to domains of prediction: economics, the nature of work and leisure, social consequences, and how men and women view themselves. These essays in prediction need to be preceded, however, by some analysis of the computer itself, and particularly its capabilities and potential in the area that is usually called artificial intelligence. This subject is taken up in the next section (*2*).

Computer Capabilities

The computer is a device endowed with powers of utmost generality for processing symbols. It is remarkable not only for its capabilities but also for the simplicity of its underlying processes and organization. Of course, from a hardware standpoint it is not simple at all but is a highly sophisticated electronic machine. The simplicity appears at the level of the elementary information processes that the hardware enables it to perform, the organization for execution and control of those processes, and the programming languages in terms of which the control of its behavior is expressed. A computer can read symbols from an external source, output symbols to an external destination, store symbols in one or more memories, copy symbols, rearrange symbols and structures of symbols, and react to symbols conditionally—that is, follow one course of action or another, depending on what symbols it finds in memory or in its input devices. The most general symbol-manipulating system that has been defined, the so-called Turing machine, requires no broader capabilities than these. The important limits on the powers of a computer are limits on the sizes of its memories and the speed of its elementary processes, and not on the generality of those processes.

There is great dispute among experts as to what the generality of the computer implies for its ability to behave intelligently. There is also dispute as to whether the computer, when it is behav-

The author is professor of computer science and psychology at the Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

ing more or less intelligently, is using processes similar to those employed by an intelligent human being, or quite different processes. The views expressed here will reflect my own experience in research with computers and my interpretation of the scientific literature. First, no limits have been discovered to the potential scope of computer intelligence that are not also limits on human intelligence. Second, the elementary processes underlying human thinking are essentially the same as the computer's elementary information processes, although modern fast computers can execute these processes more rapidly than can the human brain (3). In the past, computer memories, even in large computers, have probably not been as capacious as human memory, but the scale of available computer memories is increasing rapidly, to the point where memory size may not be much longer an effective limit on the capacity of computers to match human performance. Any estimate of the potential of the computer in the near or distant future depends on one's agreement or disagreement with these assumptions.

One common objection to the beliefs just expressed is that "computers can only do what you program them to do." That is correct. The behavior of a computer at any specific moment is completely determined by the contents of its memory and the symbols that are input to it at that moment. This does not mean that the programmer must anticipate and prescribe in the program the precise course of its behavior. A program is not a scenario; it is a strategy of action, and what actions actually transpire depends on the successive states of the machine and its inputs at each stage of the process—neither of which need be envisioned in advance either by the programmer or by the machine. A problem-solving program applied to a particular puzzle situation does not prescribe all the steps to solve that puzzle; it prescribes a selective search strategy that, when followed, may lead the computer to discover a path to a solution. But selective search, under the guidance of strategies, is also the process that people use to solve puzzles.

Of course humans, through the processes called learning, can improve their strategies by experience and instruction. By the same token, computers can be, and to some extent have been, provided with programs (strategies) for improving their own strategies. Since a computer's programs are stored in the same memories as data, it is entirely possible for programs to modify themselves—that is, to learn.

Probably the most fundamental differences between today's computers and the human information-processing system have to do with the input organs that provide the interface between the system and its environment. Simulating the capabilities of human eyes and ears has proved to be a much more difficult task than simulating the thinking processes that go on in the central nervous system. Computer capabilities in both visual and auditory domains, and particularly the former, fall far short of human capabilities.

Over the past two decades a moderate amount of work has been carried on in the field usually called artificial intelligence to explore the potentialities of the computer that have been outlined above. Some of this research has been aimed at programming computers to do things which, if done by a person, would be regarded as intelligent. Another part of the research has been directed at simulating not only the human capabilities but also the processes that human beings use in exercising these capabilities. The considerable progress that has been made in understanding the nature both of artificial and of human intelligence has hardly begun to translate itself into applications, and has been reflected to only a small degree in the actual practical uses of computers. Artificial intelligence research has had an impact upon the search algorithms that are used to solve large combinatorial problems, it is on the verge of practical application in the realm of medical diagnosis, and it has had an important influence upon certain computer programming techniques (for example, list processing). But its main significance for practical affairs lies in the future.

How, then, have computers actually been used to date? At present, computers typically spend most of their time in two main kinds of tasks: carrying out large-scale engineering and scientific calculations and keeping the financial, production, and sales records of business firms and other organizations. Although precise statistics are not available, it would be safe to estimate that 95 percent of all computing power is allocated to such jobs. Now these tasks belong to the horseless-carriage phase of computer development. That is to say, they consist in doing things rapidly and automatically that were being done slowly and by hand (or by desk calculator) in the precomputer era.

Such uses of computers do not represent new functions but only new ways of performing old functions. Of course, by greatly lowering the cost of performing them, they encourage us to undertake them on a larger scale than before. The increased analytic power provided by computers has probably encouraged engineers to design more complex structures (for example, some of the very tall new office buildings that have gone up in New York and Chicago) than they would have attempted if their analytic aids were less powerful. Moreover, by permitting more sophisticated analyses to be carried out in the design process, they have also brought about significant cost reductions in the designs themselves. In the same way, the mechanization of business record-keeping processes has facilitated the introduction of improved controls over inventories and cash flows, with resulting savings in costs. Thus, the computer not only reduces the costs of the information-processing operations that it automates but also contributes to the productivity of the activities themselves.

The remaining 5 percent of computer uses are more sophisticated. Let us consider two different ways in which a computer can assist an engineer in designing electric motors. On the one hand, the engineer can design the motor using conventional procedures, then employ the computer to analyze the prospective operation of the design—the operating temperature, efficiency, and so on. On the other hand, the engineer can provide the computer with the specifications for the motor, leaving to the computer the task of synthesizing a suitable design. In the second, but not the first, case the computer, using various heuristic search procedures, actually discovers, decides upon, and evaluates a suitable design. In the same way, the role of the computer in managing inventories need not be limited to record-keeping. The computer program may itself determine (on the basis of usage) when items should be reordered and how large the orders should be. In these and many other situations, computers can provide not only the information on which decisions are made but can themselves make the decisions. Process-control computers, in automated or semiautomated manufacturing operations, play a similar role in decision-making. Their programs are decision strategies which, as the system's variables change from moment to moment, retain control over the ongoing process.

It is the capability of the computer for solving problems and making decisions that represents its real novelty and that poses the greatest difficulties in predicting its impact upon society. An enormous amount of research and devel-

opmental activity will have to be carried out before the full practical implications of this capability will be understood and available for use. In the single generation that modern computers have been in existence, enough basic research has been done to reveal some of the fundamental mechanisms. Although one can point to a number of applications of the computer as decision-maker that are already 20 or 25 years old, development and application on a substantial scale have barely begun.

Economic Effects of Computers

The direct economic effects of introducing computers as numerical calculators and decision-makers are like those of introducing any new form of capital that raises productivity and also improves the quality of the product. The computer (its hardware together with the associated system-programming costs) represents an investment in a capital-intensive, laborsaving device that has to justify itself, in competition with other possible forms of investment, through savings in clerical and other personnel costs together with the improvements it brings about in organizational decisions.

When the main motive of introducing the computer is to mechanize existing clerical operations—in the actuarial department of an insurance firm, say, or the accounting department of a manufacturing concern—then its main economic advantage stems from the reduction in clerical costs. When it is introduced to mechanize decision processes—engineering design, for example, or control of stock or cash inventories—then its direct effect shows up as some form of productivity increase in the organization's operations. In either case, there is nothing special about the computer that distinguishes it, in its economic effects, from any other capital investment. Any such investment can be expected to have a direct effect upon employment in the organizational components where it is introduced. When the accounting system is mechanized, fewer clerks and bookkeepers are needed, else there would be no economic motivation for mechanizing. Of course, if part of the motivation for the change is to improve the quality of the system's output, the operation may be expanded, and the net reduction in personnel may be smaller than would be estimated solely from the increase in efficiency. If there is sufficient elasticity of demand for the activity, personnel may actually increase.

The most important question, however, is what the reduction in personnel at the point of impact means for the total level of employment in the economy. Again, this is a general economic issue—of technological unemployment—that does not depend on any special properties of computers. They are simply one among the many laborsaving devices that have been appearing since the beginning of the Industrial Revolution (and before).

Both standard economic analysis and a large body of empirical evidence demonstrate that there is no relation, positive or negative, between the technological sophistication of an economy and the level of employment it maintains. From a systems standpoint, a cost reduction in any part of the system releases resources that can be employed to increase the output of goods and services elsewhere in the system. At any level of employment, from 0 to 100 percent, the total revenue received by wage earners and owners of capital and land as wages, interest, and rent is just sufficient to purchase the total bundle of goods and services that is produced. Economists sometimes disagree as to why economies do not always operate at or near full employment, but they are unanimous in agreeing that the reason is not that they produce more than they can consume. (Even Marxists agree with this proposition, although they argue that full employment cannot be maintained within the institutions of capitalism.)

An even stronger statement can be made about the systems effects of cost-saving technological innovations. We usually describe devices like computers (and most other machinery) as laborsaving because they require a lower ratio of labor to capital than the methods they displace. But if we measure savings relative to output, they are usually both laborsaving and capital-saving. That is to say, a smaller capital investment per passenger mile is required to transport people by jet plane than to transport them by ox cart. Similarly, a smaller capital investment per multiplication is required if a large modern computer is used to do the arithmetic than if it is done on a desk calculator or with pencil and paper. (Do not forget to include the capital cost of the desk at which the clerk sits and the heated or air-conditioned room in which he or she works.) Now it is easy to show, for economic equilibrium and under reasonable assumptions about the supply of capital, that the introduction of capital-intensive, cost-saving innovations will raise the level of real wages and increase the fraction of the total revenue that goes to wages. This prediction from economic theory is amply supported by the histories of the industrialized economies over the present century. As productivity has increased (mainly as a consequence of technological innovation), real wages have steadily risen, as has labor's share in the total national product (4).

Now the rate of technological change depends both upon the rate of discovery of new innovations and upon the availability of capital to turn them into bricks and steel (or wire and glass). In this process, computers compete with other forms of technology for the available capital. Hence, the process of computerization is simply a part, currently an important part, of the general process of technological change. It might be described, paraphrasing Clausewitz, as "a continuation of the Industrial Revolution by other means" (5).

In taking this very global and bird's-eye view of the economics of mechanization, we should not ignore the plight of the worker who is displaced by the computer. His plight is often genuine and serious, particularly if the economy as a whole is not operating near full employment, but even if it is. Society as a whole benefits from increased productivity, but often at the expense of imposing transient costs on a few people. But the sensible response to this problem is not to eschew the benefits of change; it is rather to take institutional steps to shift the burdens of the transition from the individual to society. Fortunately, our attitudes on these questions appear to be maturing somewhat, and our institutional practices improving, so that the widespread introduction of the computer into clerical operations over the past generation has not called forth any large-scale Luddism. In fact, during the depression that we are currently experiencing, in contrast to some earlier ones, technology has not been accused as the villain.

Effects on the Nature of Work

We see that, so far as economic effects are concerned, the computer simply provides a particular path toward higher productivity through industrialization. Whatever benefits it produces, it produces in this way; whatever problems it creates, it creates as other capital-intensive innovations do. We must be careful, however, not to evaluate social change solely in terms of its impact on wages and employment. Of equal importance are the effects it may have on the

workplace, and even on leisure. Today we frequently hear the claim that computers and automation dehumanize work and that dehumanization, in turn, causes alienation from work and society. These charges have been laid not only against contemporary developments in automation but against the whole process of industrialization. They were stated eloquently in the *Communist Manifesto* more than a century ago and by numerous social critics before and since. There has been a new surge of concern with the alienation issue during the past 10 years.

Three questions need to be asked about alienation. First, how much alienation is there—is there evidence that alienation has been increased by computers and automation or, for that matter, by other forms of industrialization and mechanization? Second, in what ways is the nature of work, and the satisfactions derivable from it, changed by automation of the workplace? Third, as automation eliminates certain kinds of jobs in favor of others, what are the net effects upon the profile of jobs in the economy—are the jobs that are eliminated, on balance, more or less satisfying than the new ones that are created to replace them?

Objective data on national trends in job satisfaction are available only for about the last 20 years. About 15 national surveys have been conducted since 1958 by professional polling organizations that included questions on job satisfaction. Although 20 years is a short time, it does cover almost the whole period of the introduction of computers; hence these data should help answer the question before us. The polls provide absolutely no evidence for a decrease in job satisfactions over this period. If alienation has been increased by automation, the increase somehow does not show up in answers by workers to questions about their attitudes toward their jobs (6). Notice that the polls do not show that workers are enthusiastic about their jobs, only that they do not seem to like them any less today than they did in 1958.

Unfortunately, comparable data are not available to measure the longer trends in job satisfaction over the whole past two centuries or so of industrialization. Perhaps even if computers and automation do not intensify alienation, they confirm and complete a loss of satisfactions that was produced by the rise of the factory system. The answer to that question must be mainly speculative. Clayre, however, recently threw some interesting light on it by examining the attitudes toward work expressed in preindustrial folk literature and popular ballads (7). He finds few indications of a Golden Age in which work was generally regarded as pleasurable and satisfying. He concludes that, in general, daily work was the same burdensome necessity for peasants and craftsmen as it is for factory workers and clerks; and that life's satisfactions and pleasure were mainly sought, then as now, in leisure, not work.

Perhaps, however, we should not try to detect alienation in this indirect way but should look directly at the workplaces where computers have been introduced, in order to see how they have changed the nature of work and its environment. Sizable differences have been found in worker satisfaction among blue-collar workers in different kinds of factories, some of the important variables being job variety and worker control over the timing of work. It is not the case, however, that the most advanced forms of industrialization and automation produce the most tedious and restrictive jobs. On the contrary, those forms of work organization that appear to have been most alienating—typified by the auto assembly line or large-scale hand assembly operations—are declining in importance relative to other forms of mechanization. Blauner, for example, studied four industries in depth: printing, a traditional craft industry; textiles, a machine-tending industry; automobile assembly, highly mechanized with highly specialized jobs; and a highly automated continuous-process chemical manufacturing industry (8). He found few indications of alienation in printing and chemicals, considerably more in textiles, and most of all in automobile assembly. The industry that best typifies modern automation—chemicals—was substantially less alienating than the two that typify older kinds of mechanization.

If we look at office automation, we see that, here too, the kinds of jobs that are displaced tend to be those that are most repetitive and restrictive. Whisler, who studied about 20 companies in the insurance industry, found that computerization had produced only small and conflicting changes in the nature of clerical and supervisory jobs (9). The new jobs placed greater demands on the employees for accuracy and reliability in performance, but they were not generally perceived as being significantly more or less pleasant or more or less boring than before. And, perhaps most important of all, whatever effects were produced were small effects. Automation and computerization do not appear to change the nature of work in a fundamental way.

Again we must look not just at immediate impact but at system effects. Factory and office automation are laborsaving technologies. The jobs they eliminate are mostly jobs that were already relatively routine. Therefore, when we look at the impact on the labor force as a whole, we expect to see automation bringing about an overall decrease in the percentage of persons engaged in routine work of these kinds.

Correspondingly, there will be a larger percentage of employees than before in service occupations and probably also in technical occupations. The work-satisfaction studies discussed earlier show differences among occupational groups of precisely this kind. From these data it appears that if factory operatives and clerical workers decline as a fraction of the labor force, while service workers, sales personnel, and professional and technical workers increase, there will be a net increase in reported job satisfaction—unless, of course, a compensating shift takes place in aspiration, a possibility we must not dismiss.

On all counts, then, we must acquit the computer technology of the charges that it has been and will be a cause of widespread alienation from work. Empirically, we find no signs of a downward trend in work satisfactions, and when we look at the actual impact of automation upon the workplace and the work force, we find no reason why such a trend should be expected. On the contrary, the newer technologies may even have a modest humanizing effect on the nature of work. The notion of a Golden Age of work prior to the Industrial Revolution must also be dismissed as romanticism, unsupported by such evidence as has been examined.

Control and Privacy

The potential of computers for increasing the control of organizations or society over their members and for invading the privacy of those members has caused considerable concern. The issues are important but are too complex to be discussed in detail here. I shall therefore restrict myself to a few comments which will serve rather to illustrate this complexity than to provide definitive answers.

A first observation is that our concern here is for competitive aspects of society, the power of one individual or group relative to others. Technologies tend to be double-edged in competitive

situations, particularly when they are available to both competitors. For example, the computerization of credit information about individuals facilitates the assembly of such information from many sources, and its indefinite retention and accessibility. On the other hand, it also facilitates auditing such information to determine its sources and reliability. With appropriate legal rules of the game, an automated system can provide more reliable information than a more primitive one and can be surrounded by more effective safeguards against abuse. Some of us might prefer, for good reasons or bad, not to have our credit checked at all. But if credit checking is a function that must be performed, a strong case can be made for making it more responsible by automating it, with appropriate provision for auditing its operation.

Similarly, much has been said of the potential for embezzlement in computerized accounting systems, and cases have occurred. Embezzlement, however, was known before computers, and the computer gives auditors as well as embezzlers powerful new weapons. It is not at all clear which way the balance has been tilted.

The privacy issue has been raised most insistently with respect to the creation and maintenance of longitudinal data files that assemble information about persons from a multitude of sources. Files of this kind would be highly valuable for many kinds of economic and social research, but they are bought at too high a price if they endanger human freedom or seriously enhance the opportunities of blackmailers. While such dangers should not be ignored, it should be noted that the lack of comprehensive data files has never been the limiting barrier to the suppression of human freedom. The Watergate criminals made extensive, if unskillful, use of electronics, but no computer played a role in their conspiracy. The Nazis operated with horrifying effectiveness and thoroughness without the benefits of any kind of mechanized data processing.

Making the computer the villain in the invasion of privacy or encroachment on civil liberties simply diverts attention from the real dangers. Computer data banks can and must be given the highest degree of protection from abuse. But we must be careful, also, that we do not employ such crude methods of protection as to deprive our society of important data it needs to understand its own social processes and to analyze its problems.

Man's View of Man

Perhaps the most important question of all about the computer is what it has done and will do to man's view of himself and his place in the universe. The most heated attacks on the computer are not focused on its possible economic effects, its presumed destruction of job satisfactions, or its threats to privacy and liberty, but upon the claim that it causes people to be viewed, and to view themselves, as "machines" (*10*).

To get at the real issues, we must first put aside one verbal confusion. All of us are familiar with a wide variety of machines, most of which predated the computer. Consequently, the word "machine" carries with it many connotations: of rigidity, of simplicity, of repetitive behavior, and so on. If we call anything a machine, we implicitly attribute these characteristics to it. Hence, if a computer is a machine, it must behave rigidly, simply, and repetitively. It follows that computers cannot be programmed to behave like human beings.

The fallacy in the argument, of course, lies in supposing that, because we have applied the term "machine" to computers, computers must behave like older forms of machines. But the central significance of the computer derives from the fact that it falsifies these earlier connotations. It can, in fact, be programmed to behave flexibly, in complex ways, and not repetitively at all. We must either get rid of the connotations of the term, or stop calling computers "machines."

There is a more fundamental question behind the verbal one. It is essentially the question that was raised by Darwinism, and by the Copernican revolution centuries earlier. The question is whether the dignity of man, his sense of worth and self-respect depends upon his being something special and unique in the universe. As I have said elsewhere (*2*, p. 27):

> The definition of man's uniqueness has always formed the kernel of his cosmological and ethical systems. With Copernicus and Galileo, he ceased to be the species located at the center of the universe, attended by sun and stars. With Darwin, he ceased to be the species created and specially endowed by God with soul and reason. With Freud, he ceased to be the species whose behavior was—potentially—governable by rational mind. As we begin to produce mechanisms that think and learn, he has ceased to be the species uniquely capable of complex, intelligent manipulation of his environment.

What the computer and the progress in artificial intelligence challenge is an ethic that rests on man's apartness from the rest of nature. An alternative ethic, of course, views man as a part of nature, governed by natural law, subject to the forces of gravity and the demands of his body. The debate about artificial intelligence and the simulation of man's thinking is, in considerable part, a confrontation of these two views of man's place in the universe. It is a new chapter in the vitalism-mechanism controversy.

Issues that are logically distinct sometimes become stuck together with the glue of emotion. Several such issues arise here:

To what extent can human behavior be simulated by computer?

In what areas of work and life should the computer be programmed to augment or replace human activities?

How far should we proceed to explore the human mind by psychological research that makes use of computer simulation?

The first of these three issues will only be settled, over the years, by the success or failure of research efforts in artificial intelligence and computer simulation. Whatever our beliefs about the ultimate limits of simulation, it is clear that the current state of the art has nowhere approached those limits.

The second question will be settled anew each year by a host of individual and public decisions based on the changing computer technology, the changing economics of computer applications, and our attention to the social consequences of those applications.

The answer to the third question depends upon our attitudes toward the myths of Pandora and Prometheus. One viewpoint is that knowledge can be dangerous—there are enough historical examples—and that the attempt to arrive at a full explanation of man's ability to think might be especially dangerous. A different point of view, closer to my own, is that knowledge is power to produce new outcomes, outcomes that were not previously attainable. To what extent these outcomes will be good or bad depends on the purposes they serve, and it is not easy, in advance, to predict the good and bad uses to which any particular technology will be put. Instead, we must look back over human history and try to assess whether, on balance, man's gradual emergence from a state of ignorance about the world and about himself has been something we should celebrate or regret. To believe that knowledge is to be preferred to ignorance is to believe that the human species is capable of progress and, on balance, has progressed over the centuries. Knowledge about the

human mind can make an important contribution to that progress. It is a belief of this kind that persuades researchers in artificial intelligence that their endeavor is an important and exciting chapter in man's great intellectual adventure.

Summary

From an economic standpoint, the modern computer is simply the most recent of a long line of new technologies that increase productivity and cause a gradual shift from manufacturing to service employment. The empirical evidence provides no support for the claim sometimes made that the computer "mechanizes" and "dehumanizes" work. Perhaps the greatest significance of the computer lies in its impact on Man's view of himself. No longer accepting the geocentric view of the universe, he now begins to learn that mind, too, is a phenomenon of nature, explainable in terms of simple mechanisms. Thus the computer aids him to obey, for the first time, the ancient injunction, "Know thyself."

References and Notes

1. A few years ago, Newell and I erred in predicting that certain specific developments in artificial intelligence were going to take place "within 10 years." The fact that we were optimistic about the time scale has blinded a number of critics to the basic soundness of our characterization of the nature and directions of artificial intelligence. I shall try not to make the same mistake here of predicting that very specific things will occur at very specific times. See H. A. Simon and A. Newell, *Oper. Res.* **6**, 1 (1958).
2. For more detailed discussions of these topics, see H. A. Simon, *The New Science of Management Decision* (Prentice-Hall, Englewood Cliffs, N.J., ed. 3, 1977).
3. The position that computers can be programmed to simulate an indefinite range of human thinking processes is developed in detail, and a large body of supporting evidence is examined in A. Newell and H. A. Simon, *Human Problem Solving* (Prentice-Hall, Englewood Cliffs, N.J., 1972); and in J. R. Anderson and G. H. Bower, *Human Associative Memory* (Winston, Washington, D.C., 1973).
4. For a fuller discussion of this evidence, see (*2*, chap. 4).
5. K. von Clausewitz, *On War* (Modern Library, New York, 1943), p. 596.
6. The polls under discussion were conducted by the Survey Research Centers of the Universities of Michigan and California, the National Opinion Research Center, and the Gallup Poll. For a detailed analysis of these data, see R. P. Quinn and L. J. Shepard, *The 1972–73 Quality of Employment Survey* (Institute for Social Research, University of Michigan, Ann Arbor, 1974).
7. A. Clayre, *Work and Play* (Harper & Row, New York, 1974).
8. R. [illegible], *Alienation and Freedom: The Factory Worker and His Industry* (Univ. of Chicago Press, Chicago, 1964).
9. T. L. Whisler, *The Impact of Computers on Organizations* (Praeger, New York, 1970). I. R. Hoos [in *Automation in the Office* (Public Affairs Press, Washington, D.C., 1961)] reaches more pessimistic conclusions than Whisler about the impact of the computer, but she mainly observed transient effects at the time the new technology was being introduced, and not the longer-term effects after the changes had been digested.
10. Two books that attack artificial intelligence on this ground are H. L. Dreyfus, *What Computers Can't Do* (Harper & Row, New York, 1972) and J. Weizenbaum, *Computer Power and Human Reason* (Freeman, San Francisco, 1976). The two books have little in common except a shared antipathy against the "machine" view of human thinking, and an eloquent contempt for those who hold that view. Weizenbaum's book is the technically more competent of the two, with respect to its understanding of the current state of the computer programming art.

Trends in Computers and Computing: The Information Utility

Stuart E. Madnick

The complexity, interdependence, and rapidity of events in modern society have accelerated demands for more effective ways to store, process, and manage information. Advances in both computer hardware (electronics) and software (programming) have provided the technology that can make it possible to effectively address many of these demands. In this article I review the structure of classical computer-based information systems and then consider these advances and show how they fit into the evolution of the information utility.

Information Systems

The need for more effective information management for both decision-making and operational efficiency is being felt for many reasons, such as:

1) Scarce resources. As we have become more aware of the scarcity of natural resources and the disruptions and perturbations in their supply (such as cost increases, boycotts, and strikes), it has become necessary to be much more responsive in decision-making in both the public and the private sectors.

2) Service productivity. There has been a growing demand for human services in most of the industrialized world. More than 66 percent of the American work force is employed in providing services; this includes teachers, lawyers, accountants, bank tellers, mailmen, and office workers. As noted by Hollomon (*1*), "The U.S. has evolved into a service economy and productivity is hard to manage in service-oriented industry. . . . Information technology is the most important element in a service economy." Information systems can improve the management of complex human services and provide greater operational efficiency, for example, through electronic funds transfer, electronic mail, automated office equipment, and electronic filing.

3) Complexity of society. As society becomes more complex, it is increasingly difficult to effectively assess the subtle interdependence of many decisions. As a result, an "optimal" decision to resolve one problem may, in fact, precipitate worse problems in other areas. The frequency and significance of these counter-intuitive reactions are increasing (*2*) and this has prompted efforts to develop effective decision-support systems (*3*).

The element common to all of these concerns is information—to make decisions that can minimize the negative impacts of limited resources, to effectively manage the distribution of human services, and to cope with the complexity of society. Recent advances in computer and communications electronics, system design, and management science provide an increased potential for developing highly effective and efficient information systems.

The author is an associate professor of management science at the Alfred P. Sloan School of Management, a cofounder of the Center for Information Systems Research, and an adjunct member of the Laboratory for Computer Science (formerly Project MAC), all of the Massachusetts Institute of Technology, Cambridge 02139.

This article is based on an internal report prepared in cooperation with Dean Witter & Co., Inc.

Information Utility Evolution

As information system processing has become an increasing, if not the dominant, component of computer usage, it has become necessary to reexamine our traditional view of computer usage. In place of the "analytical engine" for numerical calculation, an information utility has evolved in which the computer is used primarily to assist in the generation, storage, retrieval, and distribution of information.

Let us consider a simple scenario illustrating the use of an information utility. On the request of the district sales manager of a large corporation, a staff member has prepared the district's sales forecast for the next 12 months. The scenario continues from that point.

1) A secretary, using a TV-like computer display terminal, types and edits the report. This process is often called computerized text processing or word processing.

2) The secretary then instructs the system to forward the report to the district sales manager. This process is called electronic mail. Note that the report remains in computerized form and is not necessarily typed on paper at any point.

3) The sales manager receives an indication on his display terminal that there is "mail" for him. When it is convenient, he examines the report on his terminal. Before sending the report out, he requests retrieval of the last few sales forecast reports to check for consistency. This process is usually called information retrieval or text retrieval. He instructs the system to electronically file the new report for future reference by its date and under "sales forecast" and the author's name. He then forwards a copy to the vice president of sales at the corporate headquarters. In existing electronic mail systems, a five- to ten-page report can be delivered, even cross-country, in a few seconds.

4) After the vice president of sales has read the report, he forwards it to a member of his staff to prepare the corporate sales forecast. The staff member extracts the sales forecast information from all the district reports and also retrieves historical and economic data available from other sources. He then uses the computer to perform quantitative analysis and evaluate certain forecast models to arrive at a corporate sales forecast that can be used for production and financial planning.

The scenario, of course, could continue, but this brief excerpt should give the reader an idea of the operation of an information utility. There currently exist embryonic versions of such an information utility in several corporations and government agencies. Considerable research and the development of the information utility industry are still necessary before such a system becomes generally available.

Research on the technical and policy issues of telecommunications and electronic mail has been reported in several recent articles (5–9). This article will focus on the trends in computer technology and structure that can support the storage and retrieval aspects of the information utility.

Information System Structure

Computer-based information systems are now being used for a wide array of applications, ranging from inventory control and production scheduling, to airline and rental car reservations, to energy policy analysis. The overall structure of a contemporary information system is depicted in Fig. 1. The basic computer system consists of the computer hardware, including the telecommunication lines and large-capacity storage devices, and the operating system software, which is the program used to control the basic hardware resources (4).

The information system software can be conceptually divided into four major components.

1) Telecommunications management system software handles communications with other computers and users at remote terminals.

2) Database management system (DBMS) software handles the general information storage and processing. It is usually the most complex component, and in many ways is the heart of the system.

3) Application-dependent software performs specialized computations, such as evaluating a specific sales forecasting model. In many information systems there may be no need for specialized application-dependent software, or it may be handled by separate "personal" computers or "intelligent" terminals, as described later.

4) Information system manager software handles the coordination between the telecommunications, database, and application-dependent software as well as the interaction with the basic computer system.

The hardware and software requirements for information systems are outlined in the following sections with specific emphasis on the DBMS component.

Key Hardware Components of a Computer System

The key hardware components of a typical conventional computer configuration that might be used to support an information system are shown in Fig. 2. The central processor fetches the machine instructions and data from the main storage (or main memory) and performs the designated operations, such as add two numbers. A sequence of instructions is called a program. The input/output (I/O) processors, operating under the direction of the central processor, handle the flow of information and commands between the main memory and the I/O devices such as card readers and printers. Communications controllers provide the interface to and controlling electronics for various forms of communication lines—for example, to remote terminals through the telephone network. Secondary storage controllers provide the interface to and controlling electronics for various secondary storage devices such as magnetic disks.

In response to the requirements for improved information management, advances in many areas of computer technology (*10*), especially microprocessors, microcomputers, large-scale integration (LSI) (*11*), memory devices (*12*), firmware (*13*), and communications (6), are making dramatically different computer hardware configurations desirable and feasible. Some advances that are particularly relevant to the evolution of the information utility are summarized below.

Microprocessors and microcomputers. A microprocessor or microprocessor unit contains the electronics of a central processor unit implemented in one integrated circuit (IC) package (or a small number of such packages). An IC package typically has an area of less than 2 square inches. The power supply and other auxiliary electronics, such as timing signals, are usually external to the microprocessor. The INTEL 8080A, currently the most popular 8-bit microprocessor, is a representative example. It primarily operates on a byte (8 bits) of information at a time (4-, 12-, and 16-bit microprocessors are also common). It has 78 instructions, can perform an add in 2 microseconds, and sells for less than $20 in quantity.

A microcomputer contains the electronics of the central processor unit plus some or all of the main storage in one or a few IC packages. The newly announced INTEL 8748, for example, is conceptually similar to the 8080A but contains the microprocessor unit, 1024 (1K) bytes of reprogrammable read-only

memory, 64 bytes of read/write memory, 27 I/O lines, a timer/counter, and most other system necessities on a single IC. The 8748 is expected to sell for less than $100 in quantity.

Using microprocessor and microcomputer technologies, inexpensive yet powerful computer systems can be developed. Digital Equipment Corporation's LSI-11, for example, is basically a microprocessor-based version of its PDP-11 16-bit minicomputer. The LSI-11, with 2K bytes of memory, more than 400 instructions, and an add time of 3.5 μsec, is available mounted on an 8½ by 10 inch circuit board with various auxiliary electronics for less than $1000 in quantity.

In contrast, the ENIAC computer, completed in 1946, had 18,000 vacuum tubes, required 15,000 square feet of floor space, weighed 30 tons, consumed 150 kilowatts of electricity, and cost $400,000, yet could only perform about 5000 additions per second and needed to be programmed by physical wire connections (*14*).

The microprocessor and microcomputer industry has been growing very rapidly. As noted in the 1976 "Microcomputer systems directory" (*15*), "Roughly 20% of the companies listed *didn't even exist* last year, while about 25% are garage-sized microcorporations."

The evolution of microprocessors and microcomputers has made it economically possible to distribute more "intelligence" throughout the computer configuration, especially in the devices, controllers, and I/O processors. Furthermore, in some cases the traditional central processor itself is being replaced by "a federation of microprocessors" (*16*), as in IBM's System/370 models 115 and 125.

As an interesting side effect of the rapidly decreasing costs of LSI processors and main memories, the costs of the other ingredients of a computer (power supplies, cooling units, cabling, cabinets, and so forth) are becoming the major ones, and are estimated to increase from about 33 percent at present to 77 percent by 1980 for conventional computers (*17*). This phenomenon makes the widespread use of processors and larger memories even more attractive economically.

Memory technologies. Traditionally, computer direct-access storage has been dominated by two fairly distinct technologies: (i) ferrite core and, later, metal oxide semiconductor LSI memories with microsecond access times and relatively high costs, and (ii) rotating magnetic media (magnetic drums and disks) with access times in the range 10 to 100 milliseconds and relatively low costs. This has led to the separation between main storage and secondary storage noted earlier.

The evolution and increasing deployment of various new memory technologies have produced a more continuous cost-performance range of storage devices, as depicted in Table 1 (*10, 18–20*). Note that these technologies, which are arbitrarily grouped into six categories, result in storage devices that span

Table 1. Range of storage devices.

Storage level	Random access time (seconds)	Sequential transfer rate ($\times 10^6$ byte/sec)	Unit capacity (bytes)	System price (cents per byte)	Technology
Cache	100×10^{-9}	100	32×10^3	50	Bipolar LSI random-access memory
Main	1×10^{-6}	16	512×10^3	10	Metal oxide semiconductor LSI random-access memory, ferrite core
Block	100×10^{-6}	8	2×10^6	2	LSI shift registers, bulk ferrite core, charge-transfer devices magnetic bubbles
Backing	2×10^{-3}	2	10×10^6	0.5	Fixed-head magnetic disks and drums, charge-transfer devices, magnetic bubbles, electron beam
Secondary	25×10^{-3}	1	100×10^6	0.02	Moving-head disks
Mass	1	1	100×10^9	0.0005	Automated tape handlers, optical (laser) beam

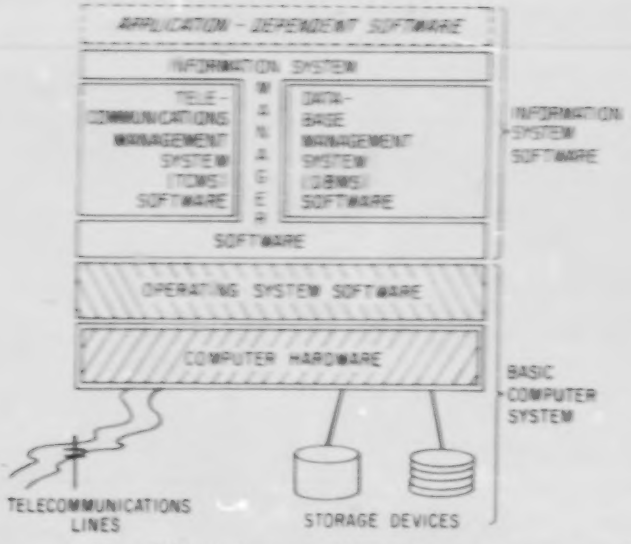


Fig. 1 (left). Contemporary information system structure.

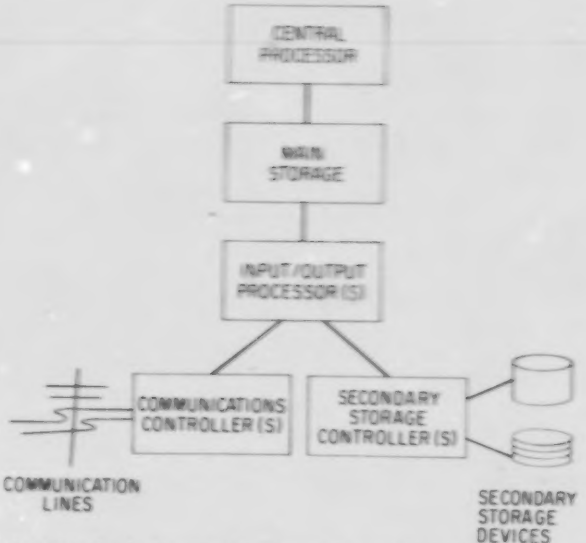


Fig. 2 (right). Typical conventional computer configuration.

Table 2. Programming languages.

A. Sample database: EMPLOYEES

NAME	OFFICE#	EXT#	DEPT#	JOB-TITLE
ADAMS, JOHN	E53-330	3-7156	15	TECHNICIAN
BROWN, TOM	39-604	3-6711	15	MANAGER
DOUGLAS, JOE	E52-154	3-2116	14	ENGINEER
...	...	...	...	...

B. Database query language example

```
PRINT NAME AND EXT# FOR ALL EMPLOYEES
WHERE DEPT# = 14 AND JOB-TITLE = 'ENGINEER';
```

C. Conventional high-level programming language equivalent

```
        OPEN FILE (EMPLOYEES) INPUT;
        ON ENDFILE (EMPLOYEES) GOTO DONE;
NEXT:   READ FILE (EMPLOYEES) INTO (EMPLOYEE);
        IF EMPLOYEE.DEPT# = 14 & EMPLOYEE.JOB-TITLE = 'ENGINEER'
        THEN PRINT (EMPLOYEE.NAME, EMPLOYEE.EXT#);
        GOTO NEXT;
DONE:   CLOSE FILE (EMPLOYEES);
```

D. Conventional low-level (assembly) language equivalent

```
        OPEN    FILE=EMPLOYEES, MODE=INPUT, ENDFILE=DONE
NEXT:   READ    FILE=EMPLOYEES, TYPE=SEQUENTIAL, AREA=EMPLOYEE
        CLC     EMPDEPT,=F'14'
        BNE     NEXT
        CLC     EMPTITL,=C'ENGINEER'
        BNE     NEXT
        MVC     OUTNAME, EMPNAME
        MVC     OUTEXT#, EMPEXT#
        WRITE   FILE=SYSPRINT, TYPE=SEQUENTIAL, AREA=OUTPUT
        B       NEXT
DONE:   CLOSE   FILE=EMPLOYEES
```

more than six orders of magnitude in both random access time (from less than 100 nanoseconds to more than 1 second) and system price per byte (from more than 50 cents to less than 0.0005 cent). This phenomenon blurs the distinction between main and secondary storage and is forcing a serious reappraisal of traditional computer architecture. There is a significant trend toward an automatically managed multilevel storage hierarchy (*18*).

Furthermore, the information in Table 1 shows that a single unit of mass storage can store and provide rapid access to the equivalent of 25 million pages of single-spaced text (assuming about 4000 characters per page) for a cost of under 2 cents per page. Coupled with anticipated further cost reductions, this provides a tremendous potential for displacing paper for many uses, leading to electronic filing and electronic mail.

Information System Software

Information systems, especially the DBMS component, provide for convenient storage, organization, restructuring, and access to information stored in the computer system.

Table 2A illustrates a simple database, named EMPLOYEES, containing employee information, in particular each employee's name, office number, telephone extension number, department number, and job title. By using a database query language, a request to see "the names and extension numbers of all engineers in department 14" can be easily expressed in a very high-level English-like manner, as depicted in Table 2B. In contrast, such a request in conventional high-level programming languages such as Cobol, Fortran, or PL/I (Table 2C) or conventional low-level assembly languages (Table 2D) would be much less user-oriented and lengthier.

This example, although it may be informative, understates the capabilities and operational requirements of most modern database systems. Instead of handling a single type of information, as in Table 2, it is often necessary to handle hundreds of different types of information, each consisting of the equivalent of hundreds of thousands of "records," as well as the complex relationships between them. For example, consider a system containing manufacturing information on parts inventory, production schedules, customers information, sales orders, and so on. A request to such a system might be, "List all customers in the northeast sales region whose current outstanding orders cannot be satisfied by our inventory on-hand or planned production within 30 days."

The services provided by most database management systems can be divided into two categories: (i) physical storage management, which is concerned with the characteristics of the storage devices, and (ii) logical storage management, which is concerned with the structure of the information, especially as seen and manipulated by the user.

Physical storage management. To economically hold large databases, it is necessary to use storage devices with a low cost per byte, such as moving head disks or tape cartridges (see Table 1). But since these devices have fairly slow access, higher-speed devices, such as main memory, are typically used to temporarily hold (buffer) active portions of the database. In this way, many requests for information can be satisfied at high speed by using the main memory buffers without needing to access the slower secondary storage. This system is called locality of reference (*21*) and is widely used in database systems. (It is similar to having a bookcase of frequently used books so that it is only occasionally necessary to go to the larger, but less convenient, public library.) Since the locus of activity in a database will usually shift in time, the physical storage management component of a DBMS must keep track of the current location of all information and move information among the various storage devices as necessary to maintain high performance.

If a variety of storage devices such as those shown in Table 1 are used, physical storage management may attain very high levels of cost-performance effectiveness but, in the process, may incur considerable system overhead. Thus, many hardware and software strategies are being investigated to increase the sophistication of physical storage management while reducing its operational overhead. Several of the major approaches are discussed later in this article.

Logical storage management. Even if there were a single storage device that was extremely fast, had enormous capacity, and was very inexpensive (which is very unlikely in the near future), DBMS logical storage management functions would still be needed. These functions are multitudinous but most can be placed into two categories: (i) performance-related and (ii) mapping-related.

Many DBMS functions are related to searching, as exemplified in Table 2. A linear top-to-bottom search, as in Table 2, C and D, is conceptually very simple but can be very inefficient for large databases. For example, even if each step of a linear search could be performed in 10 μsec, it could take more than 15 minutes to locate a specific record in a database of 100 million records. Although that is still quite fast by human standards, many database systems with currently avail-

able storage devices (such as moving-head disks with 25-msec access) can accomplish such a feat in less than 1 second. This is done through various more sophisticated searching procedures (for example, binary search or "hash coding"), usually in conjunction with auxiliary data structures (for example, alphabetically ordered indexes or cross-reference lists). As the number of uses for and users of such databases increases, it becomes necessary to handle hundreds or thousands of such queries per minute. Thus, there are many software and hardware advances directed at improving the performance of DBMS logical storage management.

Although performance is important, ease of use is critical to the future of database systems. For example, the DBMS must be able to accept a high-level nonprocedural query, such as that in Table 2B, and map it into the basic functions provided by the underlying hardware (typically at the level illustrated in Table 2D).

For the DBMS to perform these mappings automatically and correctly, considerable information *about* the information stored in the database must be maintained in various forms of data dictionaries. This may range from format information, such as how many digits are in an extension number, to data validation and editing information, such as what are allowable department numbers, to privacy and security constraints, such as prohibiting employees from changing their own salaries.

Logical storage management and especially the mapping functions have played a key role in making information systems easier and more effective to use. These features do, however, consume considerable resources. In many cases, the mapping activities may take more central processor time than the actual data retrieval or manipulation requested by the user. Thus, this is an area of considerable focus for software and hardware advances.

Specialized Information System Machines

The basic design of today's computers has remained largely unchanged for more than two decades. The components have become faster, smaller, and cheaper, but the overall structure is still primarily focused on mathematical calculation rather than the storing, retrieving, and data organizing functions that computers are actually being used for today. As Mueller (*22*) noted, "The computer industry has gone through three generations of development to perfect machines optimized for 10 percent of the workload."

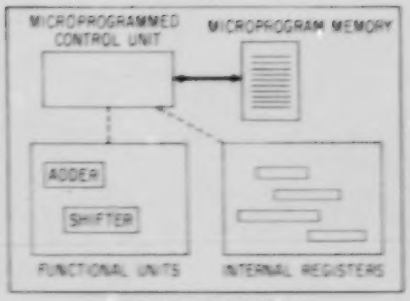


Fig. 3. Microprogrammed central processor.

The time is now ripe for a serious reexamination of the conventional computer, as exemplified by Fig. 2, especially in light of the hardware and software advances attained or forecasted. Eckert (*14*), one of the developers of the ENIAC, said, "I have often wondered why somebody didn't invent the ENIAC 10 years earlier. The tubes were almost as good then. . . . Therefore I have to really believe, at least for most of my experiences, that necessity is usually the mother of invention. . . ." It is clear that we are confronted with necessity once again; what is needed now is the imagination and leadership to apply the basic technologies to the critical problem of information management.

In this section various efforts along these lines are presented. The approaches can be largely classified into four categories: (i) firmware enhancements, (ii) "intelligent" controllers, (iii) minicomputer "back-end processors," and (iv) highly modular database machines.

Firmware enhancements. Most modern computers (as in Fig. 2) use the approach of microprogramming in the design of the central processor unit (*23*). That is, the central processor's internal registers and functional units used to decode and execute the computer's instructions are controlled by a much more primitive microprogrammed control unit carrying out sequences defined in the high-speed microprogram memory (see Fig. 3). For example, the microprogram determines what machine instruction will be interpreted to mean multiply and how the multiply operation is actually accomplished, typically by means of a number of shifts and additions. Thus, it is the contents of the microprogram memory that determines the computer's instructions as seen by a "machine language" programmer. This approach is often called microprogramming, microcoding, or firmware. By using various different microprograms, the same hardware may take on the appearance of various different computers; this is usually called emulation.

Conventional computer instructions are usually not well suited to the requirements of operating systems and database systems. Using firmware, it is possible to augment or enhance the instructions. Such an approach, called virtual machine assist, has been exploited extensively to support operating system functions in the recent IBM System/370 model 138 and 148. Similar operating system support techniques have been successfully used by many other computer manufacturers.

Firmware can be used to enhance otherwise conventional computers to make them more suitable for supporting database processing. Instead of using inefficient subroutines, each database operation may be executed as a single microprogrammed instruction. This approach has been adopted in several systems. One of the earliest efforts was part of the Lincoln Information Storage and Retrieval (LISTAR) system developed at the Massachusetts Institute of Technology (MIT) (*24*), in which several frequently used operations, such as a generalized search list operation, were incorporated into the microcode of an IBM System/360 model 67 computer. The recently announced Honeywell H60/64 has special instructions to perform data format conversion and "hashing" corresponding to frequently used subroutines of Honeywell's IDS database system (*25*). The performance advantages of this approach are highly dependent on the frequency of use of the new instructions and the extent to which they fit into the design of the overall database system software.

A more extreme approach is to define a completely new set of computer instructions through firmware. In the past, emulators have been developed that directly execute high-level languages, such as APL (*26*). The Microdata Reality System uses this approach to implement a complete database system in the microcode of an otherwise conventional minicomputer.

Since the basic hardware remains unchanged, the performance improvement achieved with firmware is largely in the higher speed of the microprogram memory (which is typically two to ten times that of main memory) and the ability of microprograms to perform certain basic operations, such as bit manipulations, very efficiently. On the other hand, since microinstructions are usually quite primi-

tive, it may take many of them to perform a specific function. With complete firmware emulation, the overall improvement in central processor performance is typically in the range of 50 to 500 percent (26)

Intelligent controllers. Many physical or logical storage management functions can be accomplished by intelligent secondary storage controllers (controllers containing microprocessors or microcomputers). For example, based on the concept of locality of reference, there have been various efforts to automate physical storage movement between devices, resulting in cache memory and virtual memory systems (*18, 27, 28*). In the IBM 3850 Mass Storage System (*29*), for example, an intelligent controller is used to automatically handle the transfer of data between high-capacity, slow tape cartridges and medium-capacity, faster moving-head disks (Fig. 4a). From the point of view of the user (that is, the central processor), the 3850 appears as a large number of disk units (Fig. 4b).

The 3850's real disks are used to hold active portions of the database, typically in units of cylinders (approximately 250,000 bytes). The controller maintains a record of the cylinders that are currently stored on the disks. When a request is received to read data from, say, virtual disk 2416 cylinder 36, the controller checks its storage location table to see if that cylinder is currently on one of the real disks; if it is, the requested data are immediately transferred to main storage. If not, the corresponding tape cartridge (for example, cartridge 4832) is fetched and scanned to locate the correct cylinder, the entire cylinder of data is then transferred to an available cylinder on one of the real disks, and the storage location tables are modified correspondingly. The requested data are then transferred from the disk to main storage as described above.

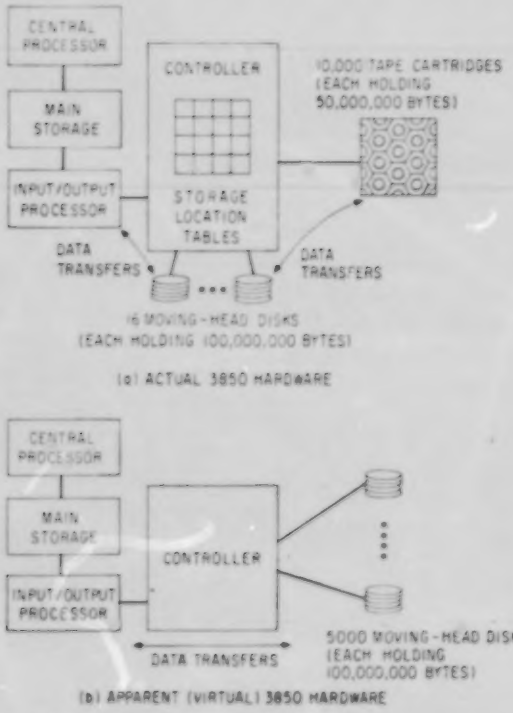
When the requested data are already on the real disks, the request can be satisfied in about 25 msec; otherwise it may take 5 to 10 seconds. Except for this possible delay, the 3850 functionally performs exactly like 5000 disks but at only 2 to 3 percent of their cost. For many applications, where only small localized portions of a very large database are used at any time, the overall performance may be almost the same.

Many storage controllers have been designed to perform certain logical storage management functions (*30*). For example, simple single-key linear search is a common feature. Linear search, even if done by the controllers rather than the central processor, is generally quite inefficient for large databases. In order to gain much higher performance, experimental controllers that perform parallel "associative" search strategies have been developed (*31–33*).

Most parallel associative search controller strategies are based on a "head per track" storage device technology (for example, magnetic drum, LSI shift registers, and magnetic bubbles) and a multitude of comparators, as depicted in Fig. 5. As each data record "rotates," either mechanically or electronically, past a read/write head, it is compared with a match record register, called the mask. If there is one comparator per head per track, as in Fig. 5, 1000 comparisons are done simultaneously; in this example a database of 160,000 100-byte records could be searched in about 10 msec, the rotation time for the storage device. In terms of conventional sequential search technologies, this is equivalent to reading and comparing one record every 60 nsec. In addition, the comparisons may be fairly complex, involving multiple data fields of the record (for example, searching for DEPT# = 14 AND JOB-TITLE = ENGINEER in Table 2).

At present, the cost of the comparators and read/write heads limits this approach to medium-size databases (10 to 100 million bytes). Furthermore, it is only well suited to storage technologies that lend themselves to low cost read/write mechanisms, and for optimal performance and operation it tends to require a fairly simple and uniform database structure.

The decline in the costs of comparator electronics, due to advances in LSI technology, makes parallel search strategies quite promising for the future. Variations on this approach are currently in use for certain real-time applications requiring



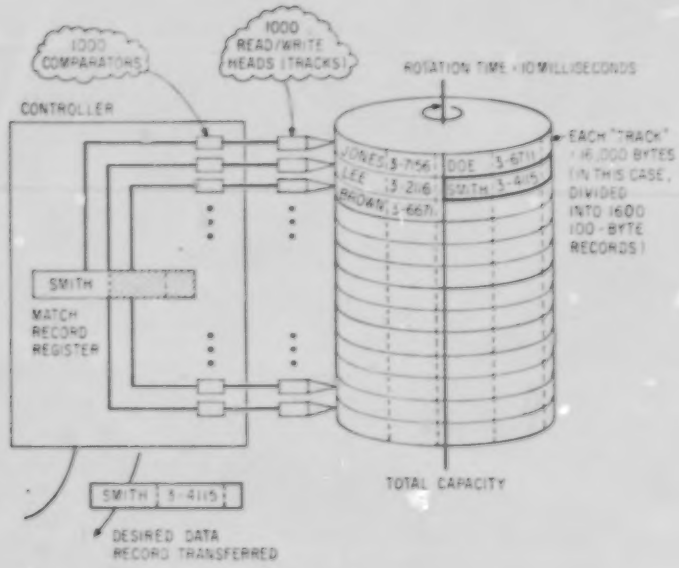


Fig. 4 (left). Use of intelligent controllers for physical storage management. Fig. 5 (right). Sample parallel associative search controller. Total capacity = 1000 tracks × 16,000 byte/track = 16,000,000 bytes (160,000 100-byte records).

high performance, such as missile tracking.

Back-end processors. As an extension of the intelligent controller approach, we can remove all of the functions of a traditional database system from the central processor and incorporate them in a back-end processor (*34*), as depicted in Fig. 6. (Front-end processor is an earlier term used to describe sophisticated communications processors.)

The back-end processor is usually a minicomputer specifically programmed to perform all (or most) of the functions of the database management system. Although such a system could be operated independent of any other computer (stand alone), we will focus our attention on the cases where the back-end processor is connected to other processors or computers, either directly to physically close systems (tightly coupled) or through a communications system to physically remote systems (loosely coupled or distributed).

There are many advantages to this approach, including the following.

1) Low processor cost. Because of the processing characteristics of database management software, a low-cost high-performance minicomputer may perform as well as (or better than) a high-cost traditional processor optimized for mathematical calculations. (For example, the ability to perform a 64-bit floating point multiply in 1 μsec is not generally needed in database processing.) By removing the database software load, the central processor can be more fruitfully used to service the remaining work load. The amount of load removable may be 40 to 50 percent or more (*35*).

2) Sharing and distributed database processing. Back-end processors can serve as shared repositories of information accessible by all of the connected computers, while at the same time enforcing defined information security constraints (*36*). The Datacomputer (*37*), operating as a loosely coupled back-end processor through the Arpanet, an international communications network, can serve as the information repository for dozens of computers. In such an environment, multiple back-end processors, each serving as a database node, may be connected to provide distributed database processing capabilities.

3) Low storage cost. By pooling the storage requirements of many computers, more economical high-volume storage devices may be used. Many of the minicomputers and specialized processors on the Arpanet use the Datacomputer because of its large capacity and low storage cost per byte.

4) Compatibility and uniformity. A large corporation or government agency may have dozens of different computers (often from different manufacturers). Interchange of information between such computers has usually been awkward because of incompatibilities between the computers and lack of uniformity in the database. By use of the back-end processor concept, the information may be conveniently shared or interchanged among these disparate computers.

Back-end processors have evolved rapidly in recent years. Some of the earliest experimental efforts include the loosely coupled Datacomputer (*37*) developed by the Computer Corporation of America using the DEC System-10 computer, and the tightly coupled XDMS (*34*) developed by Bell Laboratories by modifying the firmware of a Digital Scientific META-4 minicomputer.

The commercialization of back-end database processors is proceeding down various paths, offering different advantages to the user. As an example, Cullinane Corp., developers of the IDMS database system software for large-scale computers, is under contract with various government agencies to deliver a back-end processor, based on the PDP-11/70 minicomputer, to be attached to IBM System/360 and System/370 computers (*38*). This facility may be very attractive to users of the older System/360. Relieving the 360, often a purchased rather than rented computer, of the database processing makes additional capacity available for the increasing application-dependent production work load. This can extend the usable lifetime of the installed computer and avoid a possibly costly conversion of the production programs if it were necessary to upgrade to new computers.

Fig. 6 (left). Back-end processors, schematic.

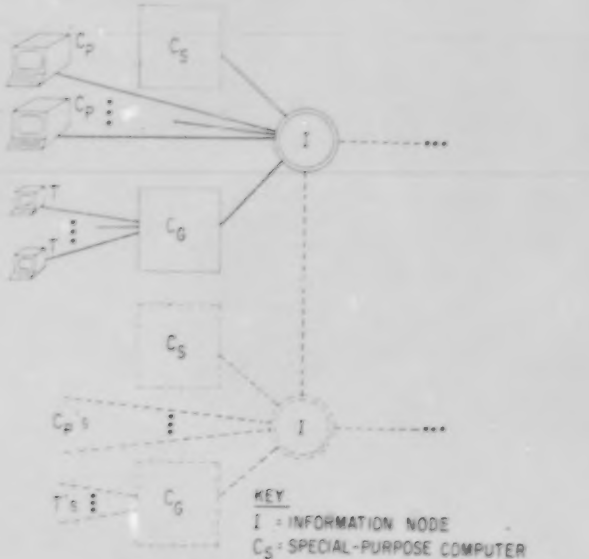


Fig. 7 (right). Information utility

It is expected that many more commercial back-end processor systems will become available in the next few years.

Database machines. The back-end processors described above were based on carefully selected and, occasionally, slightly modified conventional computers. This approach can provide significant cost-performance advantages, possibly up to tenfold, over the traditional use of the central processor for both applications programs and database processing. But there still remain major limitations on performance and reliability capabilities.

Typical contemporary information systems handle one to five requests (or transactions) per second. A system is usually viewed as having a very high volume if it needs to (and is able to) satisfy 50 to 100 requests per second. If the demand originates only from a finite number of remote terminal users (say 2000), human typing and thinking time (~ 20 seconds per request) would probably keep the load on the system within those limits. But consider the performance that would be needed for a back-end processor to serve as an information repository to a network of 100 or 1000 computers, both large-scale and personal, each capable of generating hundreds or thousands of requests per second. Such a situation—a likely one to arise in the evolution of the distributed processing and information utility concepts—could result in total request rates of more than 1 million per second. This is well beyond the capabilities of conventional computer architecture.

Despite many advances in component reliability and the use of error-checking procedures, most large-scale computer systems are very complex and have many weak spots where a failure can occur and cause the entire system to stop. A mean time between failures of 48 hours is not uncommon for many large-scale commercial computers. If the back-end processor is to service a network of users, both humans and computers, such an interruption of service could be extremely disruptive. This type of concern has led to the development of various specialized "no stop" highly redundant multiple processor systems, such as the Tandem minicomputer system and the Pluribus communications system (*39*). Conventional computer architectures are not adequate to provide the necessary degree of reliability.

Probably the most promising approach to attain both high performance and reliability is to decompose the database system into its underlying functional modules (both physical and logical storage management functions) and use a multiplicity of microprocessors to allow significant parallel processing of each function.

The DBC database computer design (*36*) uses specialized components, many based on associative search mechanisms, to handle the storage of data dictionary information, processing of directory information, storage of the database, and security enforcement.

The Infoplex design (*18, 40*), in which physical and logical storage management functions are explicitly separated, defines more than 12 separable functional levels. By using up to 4000 microprocessors, with 100 or more in parallel at each functional level, extremely high reliability and performance can be attained (all processors for a functional level are anonymous and operate under local control; if one fails, performance is reduced slightly, but operation continues uninterrupted).

Much additional research and experimentation in the design and development of database computers will be needed to meet the performance and reliability requirements of the information systems of the 1980's.

Information Utility Structure

As the evolution of the information utility continues, there will be an increased separation between computer systems specialized for the storage and processing of large amounts of information, called information nodes, and systems specialized for supporting traditional numerical computation, as well as personal computer usage. Figure 7 is a view of such an information utility. Information nodes may be interconnected, resulting in a geographically distributed database. An information node may be connected to various types of computers, as follows.

1) Special-purpose computers. These may be specialized in their construction for high-performance numerical computation (for example, a signal processor) or in their function (such as factory process control).

2) General-purpose interactive computers. These correspond to most contemporary computer systems that are used to support specific application algorithms (such as sales forecasting models) as well as program development. Users can communicate with the computer through remote terminals.

3) Personal computers. Advances in minicomputers and sophisticated terminals have spawned a movement toward distributed processing, in which calculations, word processing, and other such tasks are performed by personal computers (or intelligent terminals) rather than a more complex general-purpose interactive computer.

In all of these cases, the connected computers can draw on the information node for needed data or deposit new data, either gathered or generated. The Generalized Management Information System at MIT, for example, uses such an arrangement of virtual computers in support of the New England Management Information System (*41*) project. The purpose of the project was to develop an information system for the New England Regional Commission to assist policy-makers in managing the problems brought on by the energy crisis. Better information was needed to analyze methods to conserve fuel and to assess the impacts of tariffs and prices on different industrial sectors and states within the region. Support was also required for many other studies, such as the analysis of the merits of refineries and the impact of offshore drilling on the fishing industry. The computer-based NEEMIS project was developed to allow researchers to respond to such demands for information that had not been previously anticipated. Positive results have been reported from this project, and much can be learned from its example. Other such systems are in the development stage.

Conclusions

Various studies indicate that the installed value of computers, worldwide, will double in the next 5 years; the primary emphasis will be on a general transition from traditional batch processing to transaction processing, providing rapid response from central databases through a communications network (*42*). The evolution of the information utility concept and the accompanying technical advances are the keys to meeting these demands.

As observed thus far, information utilities will most likely be implemented first within large governmental and industrial organizations to serve their own needs. Depending on favorable governmental regulations on telecommunications and the successful implementation of the information node technologies outlined in this article, it is possible that an information utility industry, offering services to general subscribers, will be established by the mid-1980's.

Although the development of computer hardware and system software is im-

portant, it must be realized that the full potential of the information utility depends on dramatic changes in information system usage and structure (for example, electronic mail, electronic filing, automated office operation, and electronic funds transfer). In many cases, these changes are more under the control of information system designers and users than of computer manufacturers. An informed user community is an important requirement for future progress.

Summary

Demands for more effective information management, coupled with advances in computer hardware and software technology, have resulted in the emergence of the information utility concept, whereby computers specialized for information storage and processing serve as information nodes. The information nodes, which may be interconnected, can provide information management services to both conventional and personal computers. In this article the key hardware and software components of classical information systems are described to provide background on the requirements for an information utility. Four approaches to the development of specialized information nodes, drawing on various advances in technology, are presented: (i) firmware enhancement, (ii) intelligent controllers, (iii) minicomputer back-end processors, and (iv) highly modular database machines. The benefits of these advances will be systems that are more efficient, reliable, and easy to use.

References

1. H. Hollomon, *Technol. Rev.* **77**, 57 (January 1975).
2. J. W. Forrester, "Dynamics of socio-economic systems," Report D-2230-1, MIT Systems Dynamics Group, Cambridge, Mass., August 1975.
3. J. J. Donovan and S. E. Madnick, *Database* **8** (No. 9), 79 (1977).
4. S. E. Madnick and J. J. Donovan, *Operating Systems* (McGraw-Hill, New York, 1974).
5. B. Edelson, *Science* **195**, 1125 (1977).
6. D. Farber and P. Baran, *ibid.*, p. 1166.
7. M. Irwin and S. Johnson, *ibid.*, p. 1170.
8. S. E. Miller, *ibid.*, p. 1211.
9. R. J. Potter, *ibid.*, p. 1160.
10. W. Myers, *Computer* **9** (No. 11), 48 (1976).
11. R. Noyce, *Science* **195**, 1102 (1977).
12. J. Rajchman, *ibid.*, p. 1223.
13. H. D. Mills, *ibid.*, p. 1199.
14. J. P. Eckert, *Computer* **9** (No. 12), 58 (1976).
15. J. Conway and P. Snigier, *EDN* **21** (No. 21), 95 (1976).
16. H. D. Toong, *AFIPS Conf. Proc.* **44**, 567 (1975).
17. S. E. Madnick, *Technol. Rev.* **75**, 8 (July/August 1973).
18. ———, *IEEE Intercon Conf. Rec.* (1975), pp. 20/1-1 to 20/1-7.
19. R. R. Martin and H. D. Frankel, *Computer* **8** (No. 2), 24 (1975).
20. J. K. Wensley, *ibid.*, p. 30.
21. P. J. Denning, *ACM Comput. Surv.* **2** (No. 3), 153 (1970).
22. G. Mueller, *Computer* **9** (No. 12), 100 (1976).
23. S. H. Fuller, V. R. Lesser, C. G. Bell, C. H. Kaman, *IEEE Trans. Comput.* **C-25** (No. 10), 1000 (1976).
24. A. Armenti, S. Galley, R. Goldberg, J. Nolan, A. Sholl, *AFIPS Conf. Proc.* **36**, 313 (1970).
25. C. Bachman, *ibid.* **44**, 569 (1975).
26. A. Hawsitt and L. E. Lyon, *IBM Syst. J.* **15** (No. 4), 358 (1976).
27. R. L. Mattson, J. Gecsei, D. R. Slutz, I. L. Traiger, *ibid.* **9** (No. 2), 78 (1970).
28. R. M. Meade, *AFIPS Conf. Proc.* **37**, 33 (1970).
29. C. Johnson, *ibid.* **44**, 509 (1975).
30. G. R. Ahearn, Y. Dishon, R. N. Snively, *IBM J. Res. Dev.* **16** (No. 1), 11 (1972).
31. S. Y. Su and G. J. Lipovski, in *Proceedings of the International Conference on Very Large Data Bases* (Association for Computing Machinery, New York, 1975), pp. 456–472.
32. E. A. Ozkarahan, S. A. Schuster, K. C. Smith, *AFIPS Conf. Proc.* **44**, 379 (1975).
33. S. C. Lin, D. C. P. Smith, J. M. Smith, *ACM Trans. Database Syst.* **1**, 53 (March 1976).
34. R. H. Canaday, R. D. Harrison, E. L. Ivie, J. L. Ryder, L. A. Wehr, *Commun. ACM* **17** (No. 10), 575 (1974).
35. H. C. Heacox, E. S. Cosloy, J. B. Cohen, in *Proceedings of the International Conference on Very Large Data Bases* (Association for Computer Machinery, New York, 1975), pp. 511–513.
36. R. I. Baum and D. K. Hsiao, *IEEE Trans. Comput.* **C-25** (No. 12), 1254 (1976).
37. T. Marill and D. Stern, *AFIPS Conf. Proc.* **44**, 389 (1975).
38. J. Verity, *Electron. News* (3 January 1977), p. 28.
39. S. M. Ornstein, W. R. Crowther, M. F. Kraley, R. D. Bressler, A. Michel, F. E. Heart, *AFIPS Conf. Proc.* **44**, 551 (1975).
40. S. E. Madnick, *ibid.*, p. 581.
41. J. J. Donovan, *ACM Trans. Database Syst.* **4** (No. 1), 344 (1976).
42. F. G. Withington, *Datamation* **21** (No. 1), 54 (1975).

Distributive Computer Networking: Making It Work on a Regional Basis

Effective sharing through a network requires new management and resource distribution techniques.

Ronald W. Cornew and Philip M. Morse

Recent articles in *Science* have dealt with computer networking from a number of perspectives: the national overview for computer and information networks (*1*), the results of free access on an individual campus (*2*), and whether or not an academic computing center should join a network (*3*). In this article we examine the experience of establishing and operating a multisupplier or "distributive" computer network (*4*) and indicate some of the management and resource distribution techniques which the New England Regional Computing Program (NERComP) is finding useful in achieving effective resource sharing on a regional basis. Established in 1971, the NERComP network now serves educational and research computing needs at approximately 40 colleges and universities throughout the six New England states.

The NERComP network is currently unique; it is the only educationally oriented regional network involving multiple supplying institutions which is not contained within a single state or operated by a single quasi-state agency. Its success in achieving a nonsubsidized and completely cost-recovering operation appears particularly relevant to the larger task of establishing such facilities on a national basis, as discussed by Greenberger *et al.* (*1*), Massy (*3*), and others.

Dr. Cornew is director of development and former member of the board of the New England Regional Computing Program; he may be contacted at the New England Board of Higher Education, 40 Grove Street, Wellesley, Massachusetts 02181. Dr. Morse is professor emeritus of physics and director emeritus of the Operations Research Center at the Massachusetts Institute of Technology, Cambridge 02139. He has served as director of the MIT Computation Center and is the former chairman of the board of the New England Regional Computing Program.

Evolution of NERComP

The New England Regional Computing Program developed from a consortium of New England colleges and universities which, through an IBM grant beginning in 1957, was permitted to use the Model 704 computer located at the Massachusetts Institute of Technology. At that time the MIT computer represented the sole computing resource for these schools. As campus computing proliferated in the early 1960's, many of these institutions began to acquire other sources including, in a growing number of cases, on-campus comput-

ers. They continued to make use of the MIT facility, however, because it offered access to a succession of bigger, more capable machines than their own.

By 1967 it was apparent that these colleges and universities, of which more than one-half by then had their own computers, needed a variety of additional services which could not be provided by MIT alone. As they moved toward other suppliers, it became evident that a unified means of accessing the numerous large machines which were then available at academic institutions elsewhere in New England could be very useful. It was this realization that first led to consideration of a network in the region.

In order to prepare its participating institutions for the changes required in accessing a computer by utilizing a network, the consortium, with National Science Foundation assistance, engaged in a number of regional "pump priming" activities in 1968 and 1969. These included a teletype loan program in which a computer terminal and a small supplementary grant were made available to participating schools—typically smaller institutions or those lacking computer resources—to encourage them to try time sharing. The grant was intended to offset most of the costs, including computer charges and long-distance telephone toll charges, for a period of direct connection to one or more of the systems then existing in the region. For some of the medium-sized to larger institutions in New England, the consortium also participated in computer evaluation studies designed to help the institution select equipment that would meet current needs but still be consistent with future network development.

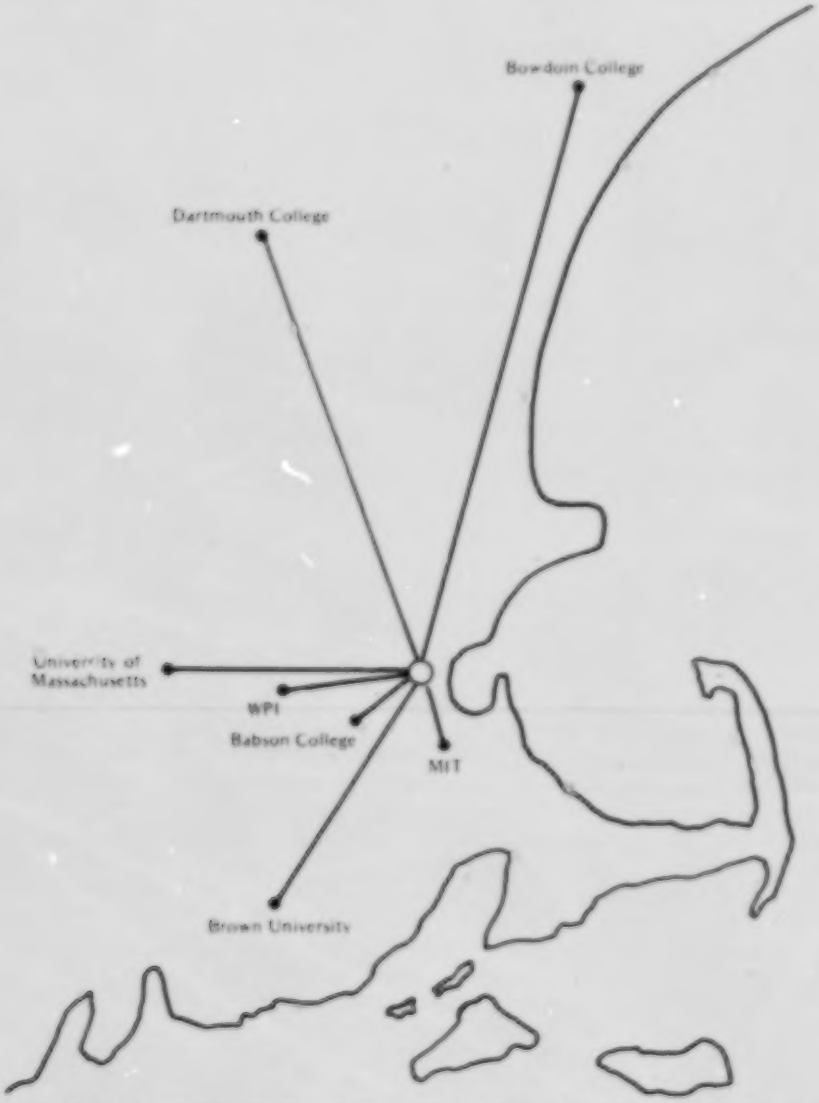


Fig. 1. The NERComP network links academic computing centers at seven New England educational institutions: Bowdoin College (PDP-10 computer), Dartmouth College (H635 dual processing system), University of Massachusetts (Cyber 74-18), Worcester Polytechnic Institute (PDP-10), Babson College (HP2000C), Brown University (System 360 Model 67), and the Massachusetts Institute of Technology (System 370 Model 168). Connections to user institutions have been omitted in this illustration.

NERComP was incorporated in 1970 as a not-for-profit corporation under the laws of Delaware, which permit formation of a corporation of corporations—a step believed necessary in order to reflect adequately the organizational nature of the consortium. Approximately 40 New England institutions of higher learning participated in the incorporation of the organization. Responsibility to its member institutions was assured through a body of institutional representatives appointed by the presidents of the dues-paying member colleges and universities. These representatives have since met at least once a year. Between general meetings, the business of the organization has been directed by a 12-member board of trustees who are elected by the institutional representatives and who serve staggered, 3-year terms. The president of the company, a salaried officer responsible for implementation of the program, was appointed to head a staff responsible for its day-to-day activities.

Current Network Facility

By 1971, while experimenting under a continuing NSF grant to determine the educational services required by users of a multicomputer network, NERComP was ready to make its first attempt at operating a network. A regional facility, centered at NERComP's office in Cambridge, Massachusetts, became operational in that year and has now linked together the seven computer centers depicted in Fig. 1.

There are four main reasons why NERComP originally chose a multisupplier or distributive network.

1) A history of cooperation in computing among the participating institutions encouraged ambitious resource sharing goals.

2) A growing number of faculty at the smaller schools with little or no exposure to computing were eager to employ the computer in their classes without having to acquire extensive programming knowledge (a step made possible by increased access to the application programs already in existence at the major computer centers in the region).

3) The growth of specialized data based systems, such as Dartmouth's IMPRESS, added impetus to the desire of individuals at member institutions of all sizes to use multiple computing facilities.

4) With long-term load sharing the discontinuous growth of individual computer

centers (necessitated by the large incremental changes in machine capabilities even within a single product line) could be smoothed by temporarily accessing remote facilities.

There never was significant interest in having NERComP establish a large central facility in an attempt to service all member needs; computing in the region was already too well developed for that and no one had any illusions that a single system could meet all needs, least of all the computer center directors who lived with the inadequacies of existing systems on a day-to-day basis.

The initial data communications network selected was a line-switching system linking users with suppliers through a switching facility located at the NERComP central office. Here the term "line-switching" refers to a telephone line connection made by physically connecting an incoming line with an appropriate outgoing line through a special switchboard or network control unit (NCU). Each user's terminal is connected to a data set (or an acoustic coupler) which is used to dial into another data set in his own locality. This second set is directly connected to a frequency division multiplexor, which in turn simultaneously transmits several data streams from other users in the same area over individual channels on a leased telephone line to a receiving multiplexor in NERComP's central switching office. There the user is connected through the central line-switcher to another multiplexor which transmits over leased lines to the supplying computer center in the same manner, with connection to this "host" computer being the ultimate result. The connection so established is bidirectional, allowing the computer to respond to the user over the same lines (see Fig. 2).

The NCU can function in either a manual or automatic mode. When a user initiates a call, his terminal is either connected to a programmed computer port, or it triggers an alarm answered by a human operator at a control terminal attached to the NCU. The operator can both send and receive messages between any user or computer and can connect the user to any available machine. Initially, users accessed the network through terminals with speeds of 10 and 30 characters per second, although facilities for higher-speed terminals, including remote job entry stations for batch processing, have subsequently been added.

By late 1973 this network had achieved a level of usage which allowed complete cost recovery. Operationally, NERComP purchased in bulk from its suppliers at wholesale rates, while distributing through the network to its member institutions at retail prices comparable to those prevailing on supplier campuses. All user billing has been handled through the NERComP central office, with the organization in turn compensating the supplying centers. While NERComP has subsequently received grant funding to subsidize development in other areas (5), including research into various forms of network management, the network operation has functioned without subsidy since that time.

A graph of NERComP's growth is shown in Fig. 3 together with a curve depicting growth of network revenue at a rate of 50 percent per year, which had been our initial anticipation. As Fig. 3 indicates, actual growth has somewhat exceeded this aim and has occurred in spite of an intervening cutback in federal funding for research and the increasingly uncertain nature of academic budgets and the economy. While break-even operation occurred at about $15,000 in monthly revenues, network income is now about twice this amount and is predicted to reach $40,000 per month when several additional services are introduced next year. NERComP has therefore achieved solely from networking revenues a scale of operation comparable to that of a medium-sized computing center. NERComP's experience demonstrates that a financially self-sufficient distributive network is possible, even on a regional basis with limited communication technology.

Operating Experience

While NERComP has purchased computing service from its suppliers on a per port basis, two types of basic service have been provided to users: connection by the hour at an incremental price, and a fixed-

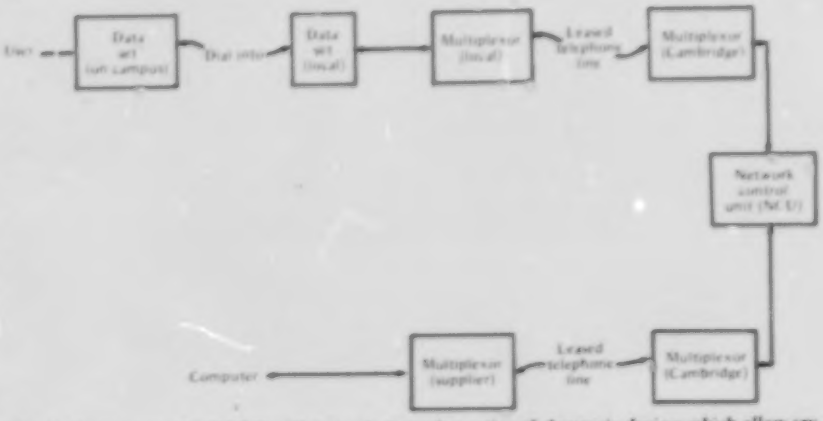
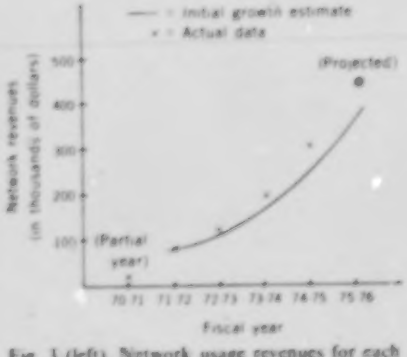


Fig. 2. Users access the NERComP network through a series of electronic devices which allow several communications to travel through leased telephone lines simultaneously. A central network control unit located in Cambridge, Massachusetts, allows the user to select the computer he wishes to utilize.



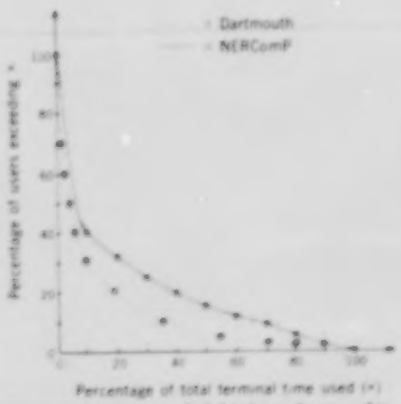


Fig. 3 (left). Network usage revenues for each year of operation since the 1970-1971 academic year. For comparison, an initial growth estimate of 50 percent per year has been included; the network has exceeded this in each year of its existence. The datum for 1970-1971 is based on a partial year, while the projected revenue for 1975-1976 relies on the 50 percent growth estimate applied to the value of the current year. Fig. 4 (right). Terminal usage of both the NERComP network (approximate) and the Dartmouth Time Sharing System. The horizontal axis indicates the percentage of total terminal time used while the vertical axis indicates the percentage of users using more than that amount. Brackets indicate the range of accuracy of the estimated values.

price monthly contract for unlimited hourly connection to a port. Of the two, the latter has proved more important to NERComP's success in reaching cost recovery operation. More than two-thirds of current income results from provision of service in this form. Hourly service was added only reluctantly and coincided with our recognition that there were two different audiences to be served by networking services. It is therefore surprising to us that some computer centers, including a number of national suppliers, have not recognized this market and included an unlimited-use option in their network offerings. In areas of heavy terminal utilization with low central processing unit demand, such as computer-aided instruction (CAI), the unlimited-use option would seem essential for reaching a sizable number of students.

In contrast to the fixed-price contract, the hourly pricing policy adopted by NERComP has become very detailed with time. Major components of service, including installation charges, modem costs, connect time, and storage, are now billed separately on a metered or as-rendered basis. With both policies, we have sought to establish rates independent of the distance between supplier and user, the current objective being rates less than $2 per hour or $200 per month per terminal on a contract basis for low-speed data transmission (10 and 30 characters per second) anywhere in the New England region. The resulting pricing policy can therefore be characterized as "cost recovery" only in the aggregate.

In reviewing the base of usage which has evolved during the 4 years of NERComP's operation, two significant and related factors emerge. Figure 4 shows the profile of network usage for a 1-month sample period in 1974. While this plot shows the customary few users consuming a large part of the resource (terminal time), there are some quantitative differences between the NERComP data and the data of its supplier, Dartmouth (*2*). For instance, 20 percent of Dartmouth's users consume about 80 percent of total terminal time, while the upper 10 percent consume approximately 65 percent of the terminal time. For NERComP approximately 60 percent of terminal time (range, 50 to 63) is consumed by 20 percent of the users, and only 30 percent (range, 26 to 34) is consumed by the upper 10 percent, which shows considerably less dedication of resources to the largest users. Thus, even if the necessary financial arrangements could be made, we are less inclined than Luehrmann and Nevison (*2*) to conclude that a policy of free access would be advisable for our network, even though it has obviously worked for Dartmouth. However, as distributive-type networking expands, the way is simultaneously cleared for individual centers to specialize in meeting the needs of their largest users.

Also of interest is the type of service these users are requiring of the network. All of NERComP's service is related to education or research. About one-half is general-purpose computing, and the remainder is largely dedicated to library programs, special packages, data bases, and CAI. There are apparently two reasons for this: (i) many users come from schools having little or no on-campus computing capability and hence must meet nearly all of their needs from the network, and (ii) where in-house facilities exist many users seek greater variety even in common services than is provided by any single center.

That this broad-based, general type of usage has developed may be surprising to those who see networks as tools for meeting specialized interests. Our experience with a regional facility suggests that an economical, distributive network will attract usage of a very general nature if this is not actively discouraged. Networks appear destined to play an unglamorous but important role in the delivery of basic academic computing service—a distribution need not well addressed in the literature.

Tax Status for Networking

NERComP evolved from computing activities initially centered at MIT and has, therefore, known both the advantages and disadvantages of being linked to a single center. In ultimately selecting a neutral location for operations together with an organization consisting of a number of equal members, NERComP was following the example of Triangle Universities Computation Center (TUCC) in North Carolina as well as that of other networking organizations (*6*). We deemed separation from any existing institution to be necessary if we were to be perceived as an independent entity able to deal with all members equally. Yet it brought with it a problem that could seriously restrict regional or national networking as a means of obtaining educational computing goals.

The Internal Revenue Service has resisted exempting consortia of tax-exempt organizations under section 501(c)(3) of the Internal Revenue Code. Thus exemption, which is important to attract the full attention and support of participating institutions and to allow the traditional benefits of freedom from taxation on purchases and income, has been obtained only through special legislation (the Common Fund, an interuniversity service organization offering advice on portfolio management to member institutions, was so exempted) and several narrow revenue rulings which have imposed severe operating restrictions.

In NERComP's case (*7*), which may set a precedent for academic computer networks, the price paid for exemption was exclusion from the network of all administrative computing—a trend recently reinforced by an unfavorable IRS ruling involving an exemption application of a consortium of state colleges in Illinois dedicated to meeting administrative computing needs (*8*).

In short, under current tax rulings educational networks could not supply administrative services such as class scheduling, billing, or processing of admissions or management information and remain exempt. Such networks might also be more generally challenged on their relatedness to the exempt educational purpose of their participants. Nor is the problem alleviated by using a (nonexempt) commercial networking carrier of the type now coming into existence because individual institutions acting as suppliers on such a network might be regarded as being engaged in unrelated business, a troublesome possibility if networking revenues become significant. Congressional relief is needed or much of the potential benefit of networking as a means for distributing academic computing resources nationally may be lost.

Networking Problems: An Overview

The use of voice-grade telephone lines employing frequency division multiplexing has limited NERComP to slow or medium-speed terminal-oriented usage. This has precluded any significant file transfer capability between computers on the network or other applications requiring computer-to-computer communications. In spite of these limitations, most members have been able to use this service as NERComP concentrated on the difficulties of networking with low-speed lines, while planning for the future. We now see that these limitations were a blessing in disguise because they allowed us to concentrate on terminal-oriented service before tackling the much more difficult problem of computer-to-computer communication. Before giving the techniques now being employed to meet this new challenge, we summarize the problems encountered in providing our original service.

1) The initial problem NERComP encountered was competition with its suppliers. By the time NERComP became operational, a number of subregional nets were already in existence supplying services from these institutions. Thus, in efforts to bring the computing services of all network suppliers into an area NERComP

usually ran into competition with the nearest supplier attempting to establish its own pattern of usage. This competition was strongest, of course, on the campus of the local supplier, where network usage could be directly controlled. But it reached well to the smaller academic institutions in the same locality, where the user was required to make a choice about who would serve his needs.

2) Allocation of computing funds between on-campus and off-campus suppliers complemented this factor in limiting initial growth. "Funny money," the term applied to departmental computer budgets which could be spent only in support of the on-campus computer center, was often a factor in making it difficult for an individual faculty member or department to purchase off-campus network services. If the institution was a NERComP supplier, this factor was an element in the competitive equation mentioned above. Research projects, which were usually independently funded, generally fared better in this regard than academic departments.

3) Network services often provided a stepping-stone to an on-campus system. No problem was more demoralizing to the staff or threatening to the early economic stability of the network than to have users disappear—often against their will—because network usage had reached the point where some kind of local computer could be justified.

4) A final problem was that network users frequently come last in the provision of services from a supplying center. This is the problem that Massy (*3*) refers to as the "organizational distance between decision-makers . . . and academic users on individual campuses." While this phase was used in reference to a computer utility, the problem is found in an even more severe form in a distributive network if one is stressing cooperative rather than competitive solutions to networking problems. It includes failure to involve remote users in local planning efforts and frequently results in the loss of users through no direct fault of the network.

In the usual single-supplier network, revenue from remote usage is usually added to an already existing base of twice as a secondary rather than primary source of income. For a distributive network such as NERComP's, attaining a critical level of usage to support its operation is, of course, a primary problem.

That NERComP was able to support the network occurred largely because it was able to identify a group of users from large and small institutions whose needs were not being met and who had control over their budgets. We have also found that institutions which acquired an in-house computer under the circumstances described under problem 3 above often came back to NERComP as their subsequent experience revealed considerable unmet needs. This has led to the realization that institutional use of a network is often oscillatory, with the network complementing the growth of local computing facilities. Networking offers a good beginning service for a growing institution and allows it to build from no utilization through several steps of machine acquisition, with remote sources supplying the unmet needs before each transition. With each successive purchase, increased dependence of individual departments on the growing central facility typically results, followed by a period of administrative relaxation in which network services are more easily procured. Through this process, however, increased utilization of networking services usually occurs until the institution eventually reaches computing maturity, having acquired the largest system it can reasonably support. It is then ready to look for the satisfaction of additional needs by both buying and selling various types of computing services through a network—that is, by "trading."

By appealing to the many smaller institutions in the region that seek computer resources, we have avoided the initial heavy reliance on striking a proper balance of trade among larger networking institutions experienced by the Michigan Educational Research Information Triad (MERIT) (*9*). Enough demand has now resulted that the larger computing institutions interested in participating in the network as both buyer and seller face an almost guaranteed favorable balance of trade. While this has helped to break down the barriers to usage of the network on their part it nonetheless remains true that the inclination of most large computer-owning institutions is to sell—a condition that NERComP has sought to work with by trying to arrange paired trades. As a recent example, a group of MIT users are obtaining Dartmouth time-sharing service in exchange for that institution's usage of the MIT facility for a number of applications involving heavy computational loads.

Network Organization: The Three Network Concept

Most of the problems that arose in the course of the network's early development, including those outlined above, were a consequence of attempts to plan for rather than with those who had a stake in the outcome. Figure 5 shows the organizational structure NERComP has adopted over the last year to address the need for broad acceptance of network concepts by all member institutions. The new NERComP organization was inspired by the EDUCOM (Educational Communications) general working seminars in 1972–1973, which distinguished three types of networks

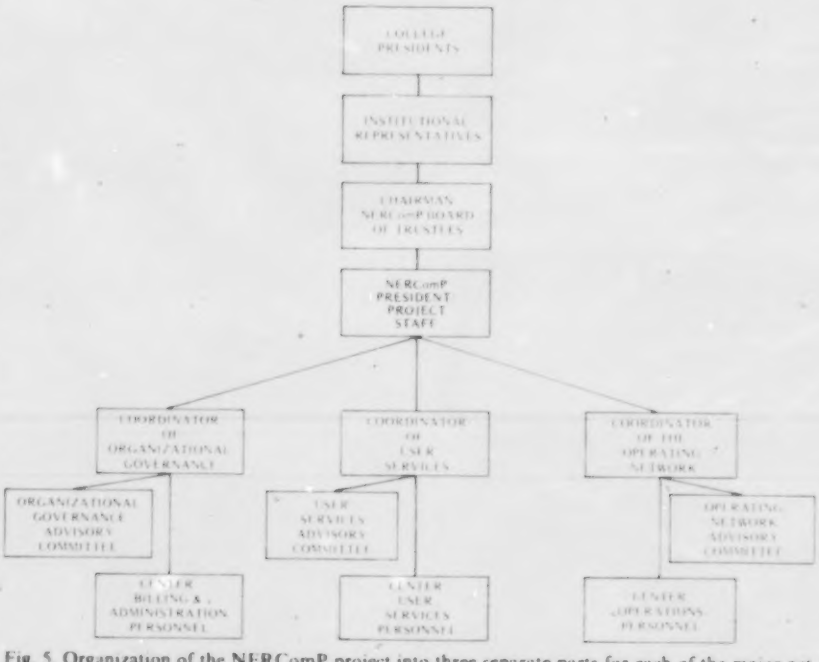


Fig. 5. Organization of the NERComP project into three separate parts for each of the major networking functions (governance, user services, and operations). Each part consists of a staff coordinator and an advisory committee made up of representatives having related responsibilities in the institutions supplying, or utilizing, the network.

within a general-purpose network (*10*). For our purposes we identify these as a governance network, a user services network, and an operating network. As shown in Fig. 5, our implementation of this concept results in an organization that differs from that described earlier by the addition of three institutional advisory committees—one for each of the subnetworks—each with a NERComP staff coordinator.

Thus NERComP is using a tripartite model not only to describe, measure, and evaluate network management concepts but also as a structure for the solution of networking problems. By adopting this network concept NERComP has selected a structure which explicitly recognizes the distributed nature of the governance and user services functions.

The Organizational Governance Advisory Committee (OGAC) is the key committee in terms of its personnel and responsibilities. Its membership includes the computer center directors at the supplying institutions, and its responsibilities include advising the board of trustees with regard to the major issues of network organization. While formally charged to develop the intersupplier arrangements related to services available, access, security, billing, and pricing policies, OGAC is emerging also as a "network market" for computer-to-computer resource sharing. The MIT-Dartmouth trade mentioned earlier resulted from these OGAC meetings, and is being monitored by this committee insofar as its organizational issues are concerned. We have found this committee to be the first in order of importance. Its early formation allows it to deal with issues which are basic and which demand resolution before other activities, such as user services, can be focused.

The User Services Advisory Committee (USAC) is composed of representatives of all the suppliers and a number of the users. Formed to develop a plan for remote distribution of user services—including training, documentation, and consultation—USAC is also developing an interesting ancillary function as a "consumers' union" where network users for the first time have the opportunity to gather collectively and present their needs to network suppliers. It is also a place where users can educate each other, and this is probably more important than any other activity in the long run. As with the OGAC, this committee also seems destined to serve as a market where buyer and seller meet. We expect that in time computer-to-computer exchanges will increasingly occupy the attention of OGAC, while NERComP's traditional terminal-to-computer usage is likely to receive substantial guidance and stimulation through the activities of USAC.

The Operating Network Advisory Committee (ONAC) is responsible for the design and operation of the physical network. While it is usually the least visible of the three committees (if everything is working correctly), it is now the center of activity for the technical design of a new message-switching network under development by NERComP (*11*). Its membership consists of leading technical people from each of the network suppliers. ONAC has responsibility for developing network interface standards and the several levels of protocols which are necessary to maintain communications among the various terminals and host computers in the network.

All of these committees are advisory to the board of trustees, which remains the major vehicle through which the representatives of the member institutions exercise formal control over the organization.

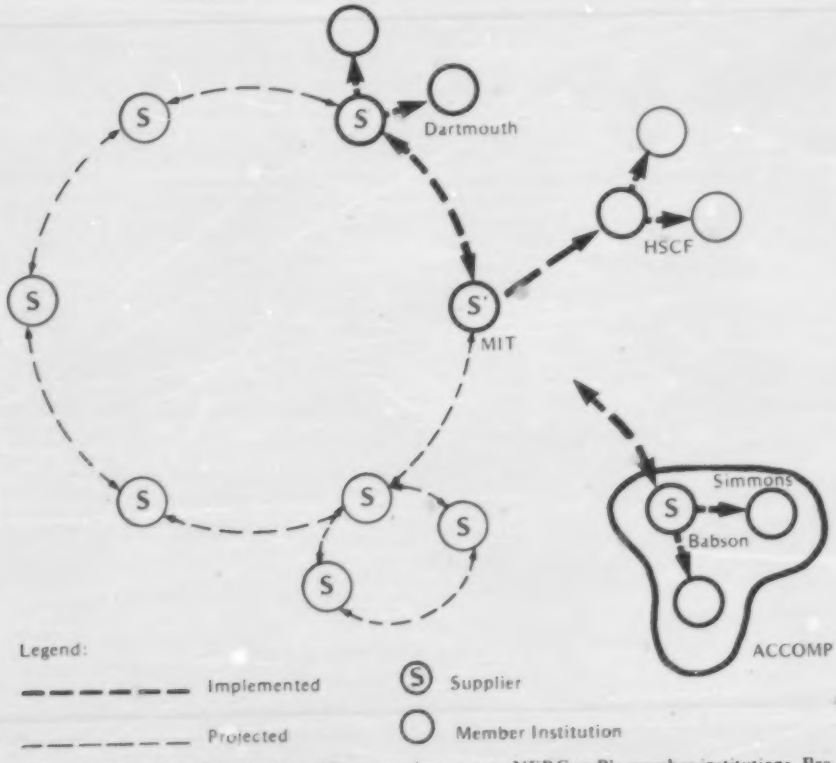


Fig. 6. Logical flow of computing and user services among NERComP's member institutions. Proceeding clockwise from the top of the dashed circle, which represents the major NERComP supplying institutions, Dartmouth is seen in its traditional role as a time-sharing supplier. Next shown is its two-way linkage to MIT, through which services are mutually distributed to users in both communities. One institution accessing the network through MIT is the Health Sciences Computing Facility (HSCF) at Harvard, which redistributes services to its medical community without itself being a network supplier. The ACCOMP group centered at Babson supplies services to its own community of colleges and is positioned to act as a secondary supplier while redistributing network services. Finally, the grouping at the bottom of the circle hypothetically shows the hierarchical concept of resource delivery evolved to an entirely new level with two redistributors with computers, each sharing services as well as supplying to the network through a nearby institution. Service connections between nonadjacent supplying institutions on the circle have been omitted for simplicity.

Resource Chaining

During much of its earlier network development, NERComP felt that it was dealing with too many problems at too great a distance and with people at too many different levels of computer sophistication. In particular, NERComP felt a growing need to decentralize both marketing of network capabilities and delivery of user services. It also became clear that a network—even a regional one—ultimately grows beyond the point where all participating institutions can relate to each other in the same way.

NERComP therefore began to explore ways to subdivide its network enterprise to maintain effective communication and to provide locally many of the services that are difficult to deliver at a distance. A view of how this might be done resulted from (i) efforts to organize a statewide network in New Hampshire; (ii) the success of the Academic Computing Group (ACCOMP), of which one of us (R.W.C.) was also a founder; and (iii) the existence of already organized networks at Dartmouth College, the

University of Massachusetts, and several other NERComP supplying institutions. These experiences taught us that certain groupings of schools (often of approximately similar size and capability) were forming naturally to meet computing needs. These groups had the advantage of small numbers, perhaps a history of planning cooperative experiments in other areas, and geographical proximity. They often provided a sufficient base to either purchase a computer or link to a network or both. ACCOMP is a case in point. Having purchased a small time-sharing system in 1970 to meet a major part of the computing needs at Simmons, Babson, Regis, Curry, and a number of other colleges in the Boston area, the consortium then began to search for ways to meet the remaining computing needs of its members. This has included continuing interest in network connections.

As NERComP learned to work with the computer centers at its major supplying institutions, it seemed natural to build on rather than to compete with these smaller groups. As it turned out, the efforts to work with each of these basically different kinds of suppliers of computing service reinforced one another, as the smaller consortia often allied themselves with a larger institution in their area to obtain access to resources not available on their own equipment, if they had any.

From these antecedents grew the concept of "resource chaining" as a framework for developing a distributive network (*12*). It offered a chance for the network to decentralize operations by providing a diverse structure through which to diffuse user services and marketing. The term itself refers to the chain of institutions through which computing resources pass from supplier to distributor to end user. In short, we found in our region that a hierarchical pattern was emerging with a number of distinct levels of participation. For the sake of simplicity we have indicated these levels as follows (*13*):

1) Regional coordinator (NERComP).

2) Area resources (Dartmouth, MIT . . .).

3) Local resources (Harvard Health Sciences Computing Facility, Babson . . .).

4) End-user institutions (Simmons . . .).

In this ordering NERComP, as the regional coordinator, provides a physical network facility while coordinating the primary network suppliers. These major suppliers, who are spread out and provide geographic coverage of the region, act as distributors of network services in their own areas of New England and are therefore referred to as area resources. They may interact with the end-user institutions either through redistributing institutions (local resources) or directly. Thus, NERComP's major suppliers are also distributors of computing services originating at other institutions; they provide a variety of different services more locally than could be provided from NERComP. The redistributor institutions may or may not possess small computers of their own, but they are counted on for even more localized redistribution of services. For any single network service the resource chain takes the form

Supplier → distributor → (redistributor) → end user

where the distributor is, in fact, another network supplier elsewhere in the region, and a redistributor may or may not be present.

A number of comments are in order. First, the resource chaining concept does allow institutions to alter their role as their ability to supply computing is altered. Second, the structure is independent of the physical network configuration, which results from reliability and cost considerations. Third, resource chaining implies that institutional computing centers exist to supply computing services and not necessarily to operate machinery. The term "information processing center" has now replaced the term "computer center" at many institutions. Perhaps an even better concept for the future is "center for the management of information processing resources," which is broad enough to cover both the operation of on-campus computer facilities, where they are cost effective, and the purchase of network services where they are not. Finally, the terms "area resource" and "local resource" refer only to the role of the institution in the distribution of resources through the network and are independent of other roles they may play. For example, both Dartmouth and MIT are each regional and national computing resources, apart from the role they may play here.

The complete distribution concept is illustrated in Fig. 6, which shows the steps NERComP is now taking, through OGAC and USAC, to reorganize its network. Such a structure, a generalization of the wholesaler-retailer concept (*14*), has some obvious consequences for charging, billing, and accessing the system:

1) Costs for a particular service to the end user will be assessed independently of where he enters the network. This requires that the supplying centers realize that service provided to a network costs them less than on-campus service, because they have reduced contact with end users. This difference provides a source of funding for both the actual cost of network communications and the remote provision of user services. Typically, 10 to 20 percent of the end user's charges will be available to the nearest distributor or redistributor for the provision of user service. (In some cases users prefer to access the network not through the nearest major node but through a more familiar institution. In these cases they will be assessed an additional charge based on the cost of providing the additional communications, but will be billed the same charge for user services.)

2) Accessing the network will also be controlled by the distributor or redistributor who is providing the service to the end user. It will be his responsibility to make sure that the user is authorized to access both the network and the computer center from which he wishes to obtain service.

3) Billing will occur from supplying center to NERComP to distributor or redistributor, with each institution adding on its appropriate value-added charge. Collecting the bill is the responsibility of the distributor or redistributor, which can limit its liability to an amount authorized in setting up the user account provided the supplier supports such a procedure. This system also requires that billing from the supplying center and NERComP occur with little delay, which is possible since most supplier institutions have computerized billing systems. Facilities are being built into the new NERComP network to expedite this cross-network billing process.

The resource chain concept can influence planning and is a model against which to compare the actual developing organization. By keeping the provision of computing service (whether generated locally or over the network) in the local area, subregional planning is encouraged, and existing relationships are strengthened.

In the same way existing regional networks, such as NERComP, may well offer a practical way to provide an intercomputer, interinstitutional national network. The alternative—educational and research networks which attempt to provide broad national service without relying on existing regional networks—would ignore a considerable body of experience (*15*) and would be likely to wind up repeating our own learning experiences. To us, national networking implies a hierarchy of resource delivery (a resource chain) with the regional network as one link: it would evidence all of the patterns of resource distribution illustrated in Fig. 6.

At the communications level the incorporation of these existing networks will require developing flexible "gateways" (*16*) to convert between the communication protocols, character sets, and message formats of whatever national carriers are adopted and those of the regional networks. This appears to be the only missing technical innovation needed for a national

network to evolve from existing regional networks.

NIMPH Development

One task of ONAC warrants special attention because of its importance to NERComP's current development effort: the Network Interface Message Processing Host (NIMPH) computer. NERComP is modifying its current line-switched network to a message-switched network. In the latter multiple communication lines are utilized between the various network nodes, and computer-originated or -destined messages (typically one line of information) are routed from one node to another to reach their ultimate destination. This is accomplished by means of codes stored in the header of the message which are recognized by a special message-processing computer at each site.

The specific factors forcing the change to a message-switched network are: (i) the poor reliability of existing communication circuits and related equipment; (ii) the cost of multiplexing and the resulting inefficient use of communication lines (less than 5 percent of the bandwidth is being used); (iii) the lack of error correction capability in the existing network, which frequently results in garbled transmissions; (iv) the need for increased flexibility in connecting users to distant computers, including the capability for a single user to make multiple network connections; and (v) the need for increased reliability through decentralization, the system should be "fail-soft" or capable of failing in such a way that local failures in lines or switching equipment have only local consequences.

Because the NERComP network includes many small institutions, it has been clear for some time that the particular message-switching technology employed in the Advanced Research Projects Agency (ARPA) network (*17*) would be too expensive for NERComP and its members unless substantial reductions in costs were to occur. As a result, NERComP has targeted a cost of $10,000 for the basic message-switching computer (NIMPH) which would be employed at a typical member installation. Initially we planned to use minicomputers as NIMPH's, but we could not find a minicomputer having the proper capabilities for $10,000. The resulting search to find a way to configure the network has led NERComP to consider microcomputers.

The design now being constructed (a central microprocessor for message switching with optional peripheral microprocessors for each of the interfaces: asynchronous for terminal handling, synchronous for line control, a host interface for connection to a network computer, and possibly interfaces to other peripheral devices such as a line printer and a card reader/punch) is modular, hierarchical, and one which can be tailored to the needs at each installation. The NIMPH incorporates in a single device host interface and message-switching and terminal-handling capabilities as in the ARPANET terminal interface message processor (TIP).

By using microcomputers to construct a cheap but reliable network operating over standard voice-grade lines (modulated to operate at 9600 bits per second), NERComP hopes to produce a low-cost message-switching system based on the emerging large-scale integration (LSI) circuit technology which could provide network capabilities in areas of the country where traffic requirements do not now warrant installation of TIP-devices and where the cost of line speeds of 50,000 bits per second or more is not justifiable.

Summary

After 4 years of operation the NERComP network is now a self-supporting success. Some of the reasons for its success are that (i) the network started small and built up utilization; (ii) the members, through monthly trustee meetings, practiced "participatory management" from the outset; (iii) unlike some networks, NERComP appealed to individual academic and research users who were terminal-oriented and who controlled their own budgets; (iv) the compactness of the New England region made it an ideal laboratory for testing networking concepts; and (v) a dedicated staff was willing to work hard in the face of considerable uncertainty. While the major problems were "political, organizational and economic" (*1*) we have found that they can be solved if the network meets real needs.

We have also found that it is difficult to proceed beyond a certain point without investing responsibility and authority in the networking organization. Conversely, there is a need to distribute some responsibilities such as marketing and user services back to the member institutions. By adopting a modest starting point and achieving limited goals the necessary trust and working relationships between institutions can be built. In our case the necessary planning has been facilitated by recognizing three distinct network functions: governance, user services, and technical operations. Separating out the three essential networking tasks and dealing with each individually through advisory committees, each with its own staff coordinator, has overcome a distracting tendency to address all issues at once. It has also provided an element of feedback between the end user and the supplier not usually present in networking activity.

The success of NERComP demonstrates that a distributive-type network can work. Our experiences in New England—which, because of its numerous colleges and universities free from domination by any single institution, is a microcosm for academic computing in the United States—indicate that such networks are best structured in a hierarchical form. This suggests that national networking should be based in part on the more than 30 existing state and regional networks (*15*).

With the groundwork now laid, we expect to see links among existing regional networks to complement development efforts now occurring at the national level. With Greenberger and others, we believe that one or more networking organizations devoted to the management issues discussed in this article will be required to facilitate resource sharing on a national scale. Because of their experience with these problems and their ability to provide service in many areas of the country through existing facilities, regional networks have a major role to play.

Afterview

The problem of motivating active institutional participation in viable cost-recovery networks is receiving substantial attention at both the regional and national levels. Knowledge concerning management of these network facilities and their benefits should grow as current experiments are completed (*18*) and use becomes more widespread. In our region we expect to see continued growth of our "traditional" network service—educational and research computing—complemented by increased networking among libraries, through organizations such as the New England Library Information Network (NELINET) and the Northeast Academic Science Information Center (NASIC), which provide online library services including cataloging and bibliographic search and retrieval assistance (*19*). We also expect that computer networks will play an increasingly important role in hospital and community health center settings, primarily to help meet the need for better continuing education facilities through computer-aided instruction.

The growth of networking has now produced a need for increased cooperation

among networks, whether computational or informational in character, or regional or national in scope. Examples occur wherever computer and informational networks operate in the same area, as well as where links between existing regional networks or between regional and national facilities are contemplated.

As with earlier "intranetwork" development involving institutional-network relationships, issues in the areas of governance, user services, and technical operations need study for networks to cooperate effectively. For instance, how do two networks sharing the same physical resource manage it for the common good? This occurs when, say, a computer network and an informational network in the same geographic area seek to use the same communications facility. What new problems arise when the distribution chain employed in delivering a service through a regional network is extended to a national level? What communication compatibility problems arise when different computer networks attempt to "talk" to one another?

This "internetwork" problem, or set of issues which must be faced in forming network-network connections, bears considerable resemblance to the intranetwork problem. While some serious exploration has begun (*20*), a coordinated program to understand the principles of internetwork cooperation remains in the future.

References and Notes

1. M. Greenberger, J. Aronofsky, J. L. McKenney, W. F. Massey, *Science* **182**, 29 (1973).
2. A. W. Luehrmann and J. M. Nevison, *ibid.* **184**, 957 (1974).
3. W. F. Massey, *ibid.* **186**, 414 (1974).
4. The terms multisupplier, multinuclear, ring, and distributive have all been used to describe the type of network under discussion here. Similarly the terms single-supplier, nuclear, star, computer utility, or more simply time-sharing network all appear in reference to a communications network supplier's from a single computing center. We prefer multisupplier or the term distributive network as used by Massey (*3*). The use of the terms ring and star in this context is, in our judgment, unfortunate, as these terms relate to network topology rather than the breadth of the services provided and encounter the further ambiguity that a distributive network may employ a starlike topology (see NERComP's current network depicted in Fig. 1).
5. The NERComP's research and development activities are currently funded by the Control Data Corporation and the National Science Foundation.
6. I. W. Cotton, *Natl. Bur. Stand. Tech. Note* 805 (1974), p. 34.
7. Revenue Ruling 74-614, *Intern. Revenue Bull. No.* 52 (1974).
8. "State-college consortium denied tax exemption," *Chron. Higher Educ.* (16 December 1974), p. 7.
9. B. Herzog, *Campus* **73**, 11 (1973).
10. M. Greenberger, J. Aronofsky, J. L. McKenney, W. F. Massey, Eds., *Networks for Research and Education: Sharing Computer and Information Resources Nationwide* (MIT Press, Cambridge, Mass., 1974). This reference also includes a discussion of the NERComP network organization at an earlier point in time. See T. E. Kurtz, "The NERComP network," p. 282. NERComP's exploration of governance, user services, and operating network structures has been undertaken with the assistance of the National Science Foundation.
11. The term "packet-switching" is often used to describe a network of the type under discussion here in which information is transmitted in blocks of limited size with origin and destination address codes included to control forwarding of the block to the proper host computer and the return of the appropriate response. We prefer the generic term "message-switching" in this context, which leaves open the question of how the message is "packetized" or broken down into smaller units for transmission. For an excellent discussion of these terms see the Final Report of the Ad Hoc Group on Packet Switching, "Study of areas for standardization in packet switching," American National Standards Institute (ANSI) Task Group X3S37 Document 75-10, 7 February 1975. I. W. Cotton, Chairman.
12. R. Ricard, former institutional representative from the University of New Hampshire and current chairman of OGAC, is responsible for the resource-chaining concept as applied to computer networks.
13. The NERC's used here were inspired by those used by the Regional Medical Library Program for the distribution of medical materials through the interlibrary loan system, administered by the National Library of Medicine. See H. M. Schoolman, *Bull. Med. Libr. Assoc.* **60** (No. 2), 284 (1972); V. Pings, *ibid.*, p. 274.
14. D. L. Grobstein and R. P. Uhlig, *AFIPS (Am. Fed. Inf. Process. Soc.) Fall Joint Comput. Conf. Proc.* **41**, 889 (1972).
15. F. W. Weingarten, N. R. Nielsen, J. R. Whiteley, G. P. Weeg, *Study of Regional Computing Networks* (Univ. of Iowa Press, Iowa City, 1973); C. J. Mosmann, *Statewide Computing Systems: Coordinating Academic Computer Planning* (Dekker, New York, 1975).
16. G. Marks of the Institute of Social Research, University of Michigan, first called this term to their attention.
17. L. G. Roberts and B. D. Wessler, *AFIPS (Am. Fed. Inf. Process. Soc.) Spring Joint Comput. Conf. Proc.* **36**, 543 (1970).
18. EDUCOM has planned a simulation and gaming project in which key administrators at approximately 15 leading academic institutions will participate in the formation of a national network model, then play a game designed to show the probable effects on their institutions over a period of years of various decisions regarding network participation. This program complements many studies carried out by existing regional consortia, including NERComP, over a number of years.
19. Both NELINET and NASIC are programs of the New England Board of Higher Education.
20. The Council for Computerized Library Networks, consisting of approximately 15 leading organizations in the area of application of computers to library networking, has recently formed to "identify, discuss and coordinate solutions to common problems." Its current interests include the internetwork governance issue.
21. The NERComP board of trustees consists of: Thomas E. Kurtz (Chairman of the Board), Director, Kiewit Computation Center, Dartmouth College; Philip M. Morse (Vice-Chairman of the Board), Director, Operations Research Center, Massachusetts Institute of Technology; John Alman (Secretary), Director, Computation Center, Boston University; Edgar T. Canty, Director, Computation Center, Babson College; Alan D. Ferguson, Executive Director, New England Board of Higher Education; Greydon C. Freeman, Director, Computer Center, Yale University; Walter Freiberger, Director, Center for Computer and Information Sciences, Brown University; Jeremy E. Johnson, Director, Computing and Data Processing Services, University of Maine; Norman Johnson, Director, Academic Computer Facility, Wheaton College; Raymond K. Neff, Director, Health Sciences Computing Facility, Harvard University; Roderick Ricard, Institutional Representative, University of New Hampshire; and Conrad Wogrin, Director, Research Computing Center, University of Massachusetts. Robert A. Rollo currently serves as President.

Microprocessors?—An End User's View

Raymond E. Dessy

There is little doubt that microprocessors are going to change the way instruments in research laboratories and analytical service areas are designed and operate and how they will interact with their operators. Within 5 years most new equipment will be using microprocessors to acquire analytical data, perform small manipulations on the data base, and report the results. Thus, it is important that microprocessors be placed in a proper perspective, especially since there is an inclination to view them as the focal point of an entirely new capability.

The 1960's saw the evolution of the minicomputer from a questionable relative of the large computer system into a strong competitor. Minicomputer systems are now capable of supporting large numbers of high-speed peripherals. They have an architecture that allows them to run sophisticated software systems under the control of programs called Executives and permits the user access to a variety of high-level languages such as Fortran, Basic, and Algol. In these languages the programmer can elicit complex mathematical operations with single-line commands. The tedious and expensive operation of programming minicomputers at the machine code or assembly language level (where each machine operation must be coded into a number that is interpreted by the computer) has disappeared, except for cases requiring speed and flexibility.

The evolution of the minicomputer was a result of the interactions of a growing market, lowered prices, lowered manufacturing costs, and the ability to put more functional electronic components (such as transistors) into a fixed amount of space. The last factor is best illustrated by a chart showing the number of components per square inch for several years since 1960 and the technological developments that permitted these advances (see Table 1).

The impetus for some of this development was a consumer market oriented toward hand-held calculators. These devices provide a large amount of computing power at slow speeds. By allowing key-stroke programming they offer the user liberation from the strictures that higher-level languages impose. However, it must be realized that they are not suitable for applications requiring speed, massive memory, or input/output (I/O) to external devices.

The high-density large-scale integrated circuit (LSI) technology also attracted the interest of equipment designers, who realized that the point-of-sale market was in need of very simple computer-type equipment. Supermarkets, gas stations, and other consumer areas needed devices suited to an accounting, inventory, and data communications environment. Computing power and speed were not important, but I/O capability was a necessity. Beginning in 1970 this need led to the availability of integrated circuits containing some of the elements and capabilities of the central processor of a computer. These were termed microprocessors. Their intended markets led to a design for handling data words containing 4 or 8 bits (*binary digits*) of information, since data communications and decimal arithmetic, which use binary coded decimal (BCD) notation, are oriented to these word lengths (BCD is a numbering code in which each decimal digit is represented by 4 bits).

At that time design and construction costs for instrument and control equipment were soaring, and microprocessors offered an interesting alternative to the hardwired logic (logic implemented by fixed, single-purpose circuits) currently being used. Automobile makers, washing machine manufacturers, and process control designers saw the microprocessor as ideally suited to the sensing, control, and reporting operations in the units they manufactured in large quantities.

To the instrument maker, microprocessors were and are viewed as a means of reducing the design costs of instruments and making them easier to manufacture and simpler (but not necessarily cheaper) to repair. The result is that numerous instruments are now being sold in which data acquisition and reporting is done by microprocessors. In many units a microprocessor also controls instrument operation, either alone or with operator interaction. Typical examples include a sophisticated electrochemical instrument, x-ray fluorescence equipment, and a variety of chromatographic units. The promise of complete instrument self-calibration and optimization and the potential to change instrument function drastically by simple changes in the operating program has been voiced. These capabilities, however, have yet to be demonstrated.

The availability of microprocessors has excited much interest in the hobby and scientific realm because of the suggestion that they provide cheap computer power and will give minicomputer capabilities at hand-held calculator prices. But microprocessors vary in the functions they perform easily and effectively, depending on their architecture—much more so than do the spectrum of minicomputers available. There are serious pitfalls for the novice who mismatches his application and his microprocessor solution.

Because of a lack of knowledge of computer operation, and particularly of software generation, the boundaries are confused between the two areas microprocessors will affect most drastically. On the one hand are applications in which control is of utmost importance and the limited vocabulary (instruction set) and computing power entailed by an 8-bit word is of no importance—the controller area. This area is in an extreme state of flux, and will remain so until production economies and competition can no longer lead to price reductions. On the other hand, during the next 2 years high-level minicomputer capabilities will be attained with sophisticated microprocessors. In the interim inexpensive microprocessor-based minicomputers will be attractive to those wanting computing power and speed for systems that will not be required to expand indefinitely in size and power—the stand-alone computer area. Both vendors and consumers have confused these two areas either by intent or by lack of understanding.

The root of the problem is that although new techniques are not being introduced with this technology—we are not doing anything new, we are doing it in a different way—the devices do require something most scientists do not

The author is professor of chemistry at Virginia Polytechnic Institute and State University, Blacksburg 24061.

fully understand or appreciate, software. To perform any task, microprocessors must be programmed. And the program must take cognizance of how data are collected, how rapidly they must be acquired, and in what ways they must be manipulated. All of this takes place in the digital environment and involves mathematical and logical operations which must currently be invoked by programmed instructions in assembly language or machine language. Program generation at this level is tedious, time consuming, and costly under the best of conditions, let alone in the environment surrounding most microprocessors. To utilize a microprocessor intended for the controller area in a task requiring one oriented to the computer realm is to invite software catastrophe. To use a microprocessor that is ideal as a standalone computer in the control environment is financially unsound.

It is my purpose in this article to present sufficient background philosophy, language, and examples that an interested user can make his own decision properly. The article is based on the conservative application principles that an instrument automation research group has developed through extended experience with both minicomputers and microprocessors.

Current Microprocessor Design

Let us examine some typical microprocessors currently being vended. This will provide a real-world framework for introducing the jargon and philosophy of the area. Four separate processors will be discussed, each one more complex and powerful than its predecessor. Although many microprocessors are advertised, relatively few are easily available in unit quantities to the interested end user. The examples given here were chosen with such availability in mind.

A microprocessor is defined, for the purpose of this article, as a set of one to four integrated circuit chips, based on LSI technology, which will provide the functions characteristic of the central processing unit (CPU) of a computer (Fig. 1). To convert a CPU into a useful computer the first step is to surround it with memory in which programs and data can be stored. Memory may be of the type where information can be both stored (written) and retrieved (read) in a random manner; this is called random access memory (RAM) or read/write memory. Where only executing programs are involved, read only memory (ROM) may be used (Fig. 2). Also needed are circuits to perform the tasks of (i) selecting a device for activation (device decoder), (ii) selecting a function the device is to perform (function control lines), and (iii) accepting and acknowledging interrupts (interrupt request and acknowledge lines). These aspects will be discussed later.

Table 1. Number of functional electronic components per square inch of space for five dates between 1960 and 1974.

Year	Components per square inch	Technological development
1960	4	Discrete elements (transistors, resistors)
1962	40	Integrated circuits
1965	400	Medium-scale integrated circuits
1969	40,000	Large-scale integrated circuits
1974	400,000	High-density large-scale integrated circuits

The CPU must have a way, or ways, of addressing memory to fetch the next instruction for execution, or fetch the piece of data for manipulation. The methods by which this can be accomplished are classed as memory addressing. As data are manipulated they will have to be stored in registers in the CPU. If arithmetic or logical operations are required, special registers to handle such tasks are needed, and these are called accumulators.

Finally, it is useful to have some simple means of establishing certain facts about the result of operations just performed in the accumulator—for example, the addition of two numbers. Such information is stored in 1-bit (yes, no) registers called flags, as indicated by the following examples.

Has a zero been left in the accumulator?
Zero flag: (1, 0)

Is the answer negative?
Sign flag: (1, 0)

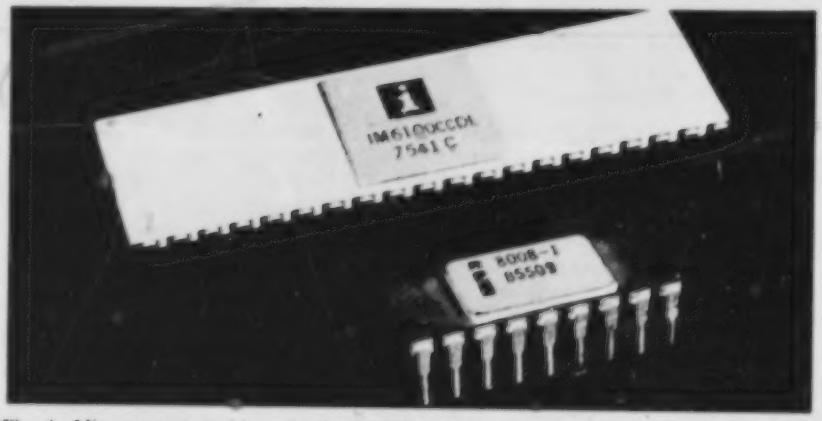


Fig. 1. Microprocessor chips: the Intel 8008 and the Intersil LSI version of the Digital Equipment Corp. PDP 8 computer.

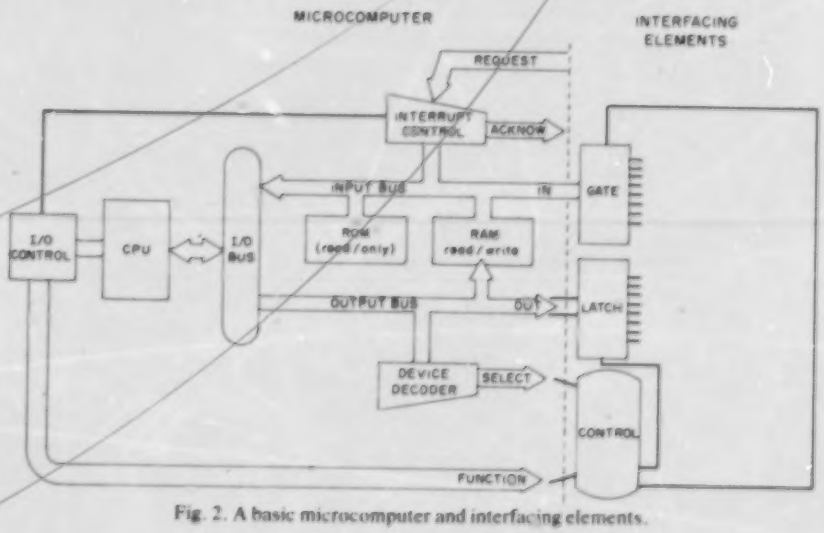


Fig. 2. A basic microcomputer and interfacing elements.

Has a carry occurred?
Carry flag: (1, 0)

Has an overflow occurred?
Overflow flag: (1, 0)

(Overflow occurs when two positive numbers summed together give a result that is considered negative by the CPU.)

All that remains is to have some means of starting our computer up and controlling the flow of the program. This is done by setting up a register called the program counter, which contains the address (number indicating a specific location in CPU memory) of the first executable instruction, and then performing the fetch operation. At this point the program counter is set to point to the next executable instruction in memory; the program counter is an arrow that always points to the next step in the program.

These features are shown in Fig. 3 for the Intel 8080 CPU. This CPU has a single accumulator in which arithmetic and logic operations can be performed. The word length for this processor is 8 bits. Data having a range of 0 to 255 in decimal notation can be accommodated. However, 8-bit-wide instructions do not permit a highly involved mathematics and control language. The situation is equivalent to limiting your own vocabulary to 8-letter words. It is also evident that since each memory address must be a unique number, 8 bits of information could uniquely identify only 256 addresses. Therefore two 8-bit-wide registers in the CPU are used to form a 16-bit-wide pointer-to-memory, thus allowing the CPU to address up to 64 kilobytes of memory (a kilobyte is 1024 words, each 8 bits wide). The operation of accessing an 8-bit data word from memory requires the following operations.

1) Load pointer-to-memory with high 8 bits of address.

2) Load pointer-to-memory with low 8 bits of address. These numbers must be part of the executing program.

3) Perform fetch from memory at the location where pointer-to-memory is set and put data into accumulator.

The Intel 8080 CPU also has four 8-bit-wide registers that can temporarily store data. These can operate independently as R1, R2, R3, and R4. A limited set of operations treat these as 16-bit-wide register pairs, (R1, R2) and (R3, R4). Zero, negative, and carry flags are provided to ascertain the result of arithmetic and logic operations that occur in the accumulator. As in most microprocessors, these consist of addition, subtraction, increment, decrement, comparison, and the simpler logic operations "and," "or," and "exclusive or."

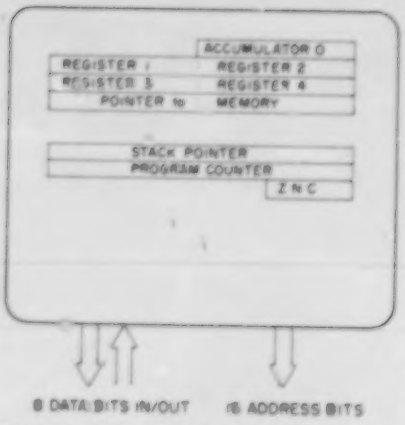


Fig. 3. Architecture of the Intel 8080 CPU.

The last feature of the hardware shown, the stack pointer, is associated with the fact that most operations in the real world are discontinuous. In daily life, as you are interrupted by a visitor while filling out a form, the interrupted task will be put aside in a mental stack where it can be picked up at a later time. While you are taking care of the visitor the telephone rings, and the visitor's problem is also placed in the stack until you hang up the telephone. Then you can go back to the visitor's problem at the point where you left him, and when he leaves you can go back to the paper work.

A computer can be designed to operate in the same way, and the program counter is the key to the whole operation. A program can be interrupted in several ways. For example, if a software request is made to jump to an area of memory where you can calculate a commonly used function (such as $\log x$), and then return when done to the executing program, you jump to subroutine. Or, to service a request by an external device for information and then return when done to the executing program, you interrupt. Each time this occurs, if you can store the program counter and the flags, as well as the contents of the accumulator and all registers, in some sort of stack for retrieval at a later date, a satisfactory operation will be assured. After each such interruption the material stored on the stack will be popped off and the program counter, flags, accumulator, and registers restored to where they were at the time the interruption occurred.

If all of this material is stored in memory the length of the stack can be unlimited, and many nested interruptions can be serviced. The function of the stack pointer is to keep track of the last information placed on the stack. When the current interrupting task is serviced the stack pointer allows the program to restore the CPU for the next shelved task, and the stack pointer is moved to show the location of the information associated with any previously interrupted operations. It should be noted that the stack operates as a first-in, last-out device, in the sense that as many successive interruptions occur the first thing placed in the stack is the last thing removed.

Two other features of the Intel 8080 require comment in order to set the stage for the discussion of the other processors. Since the restrictions on memory access imposed by the use of a pointer-to-memory register are severe, a means of loading constants and other numbers into the registers quickly and conveniently is provided. Rather than having to load the pointer-to-memory register and then execute the fetch, the CPU has the ability to address memory relative to the current value in the program counter. The concept involves loading the register with the contents of memory pointed to by the program counter, which fetches the contents of the memory location immediately after the location in which the executing instruction is found. This particular form of program counter relative addressing is called immediate mode addressing.

The second feature involves the fact that most computers manipulate numbers in purely binary fashion. The outside world often manipulates numbers in a decimal format, coded in binary, or BCD.

The Intel 8080 was originally designed for the point-of-sale market, and it does have the capability of taking two 8-bit numbers, each representing two BCD digits, adding them, and correcting the result to proper BCD notation. This is termed BCD-decimal adjust.

Application 1

A common automation task in the psychology laboratory is to provide a set of cue signals to a test animal, and reward correct responses with food. For example, a specific light, or a certain sequence of lights, being turned on should elicite a response that consists of depressing a particular microswitch. These functions are easily activated and sensed by an Intel 8080 processor. Stimulus and response data, at this level, require little mathematics. Equipped with an appropriate electromechanical interface, the microprocessor can produce the final data in a test on punched paper tape for submission to a larger computer for statistical analysis.

Very simple programs can calculate the digital numbers necessary to generate the cue signals, and these can be output by the processor. Most vendors now provide an interface board that will output such digital signals on command; for example: 1, light on; 0, light off (see latches in Fig. 2). Normally, all that needs to be added is driver circuitry to provide the voltages and currents necessary to activate the external devices. Input of information from microswitches involves digital information—Is it open or closed, 1 or 0?—and can be accomplished with vendor-supplied digital input interface boards (see gates in Fig. 2).

The fact that the stimulus and response pattern can be altered by changes in the software (operating program) rather than by hardware changes is an attractive one. Most laboratory scientists will be more attracted to learning the rudiments of programming than to acquiring expert knowledge in electronics. However, with microprocessors the programming language will be machine or assembly code.

At this point it is necessary to bring up the concept of analog and digital data. In the example above the input and output data are in digital form—all conditions can be represented by an appropriate series of 1's and 0's. The digital world comes in discrete steps, like integral numbers. Yet the world we live in is composed largely of signals in the analog domain which have an infinite number of allowed states; witness the positions allowed on the meter in most voltmeters. A computer can consume only digital data. For this reason there must be a mechanism to convert analog information, which originates in most of our analytical equipment, into digital data. This involves analog-to-digital converters (ADC) (Fig. 4). On the other hand, after a computer has digested the input data and evaluated them on the basis of programs it contains, it is often necessary for it to communicate with the outside world. If this outside world is an analog one (like oscilloscopes and plotters) it is necessary to convert a string of digital bits into an analog signal. Digital-to-analog converters (DAC) perform this task. These more complex automation techniques are used in the application discussed next.

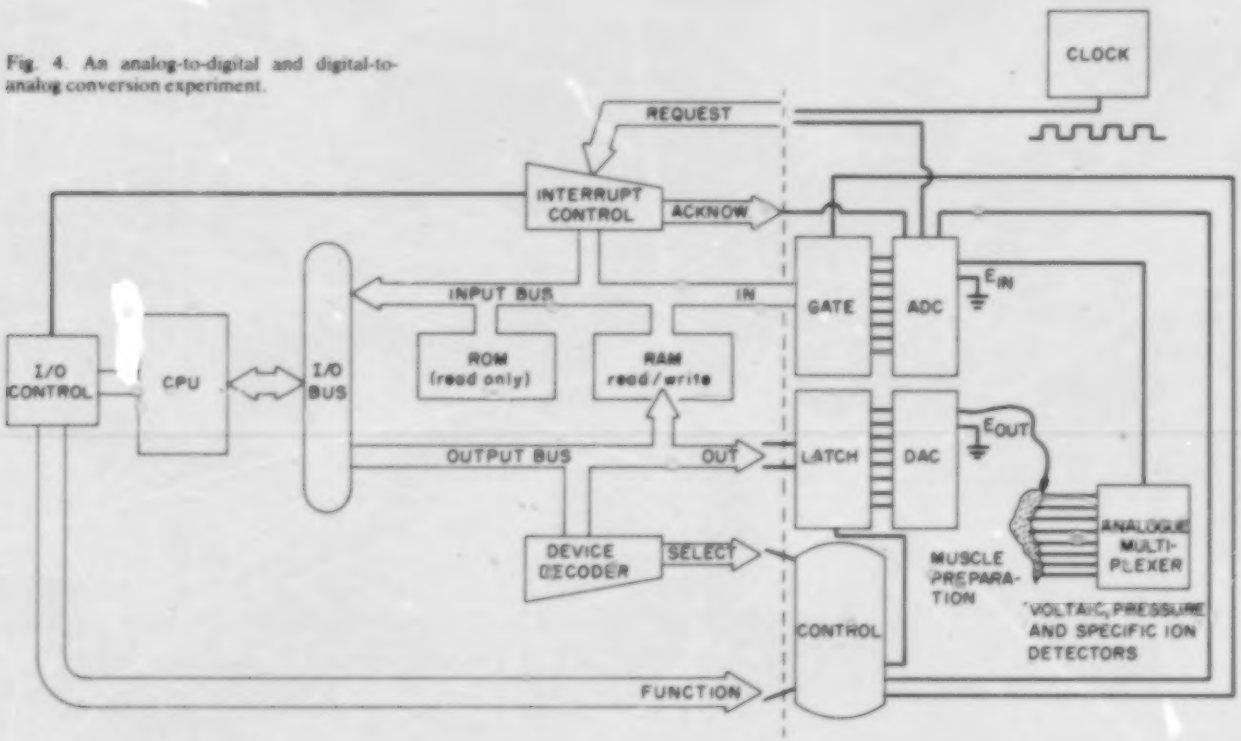
Application 2

In a biological monitoring system, fish are held in flow-through tubes and are subjected to input streams with varying amounts of contamination (Fig. 5). The respiration rate and "cough" rate can be ascertained by monitoring the inhalation-exhalation cycle with a pressure transducer. The output of the transducer is converted to digital data by an ADC. Peaks, valleys, and intensities are detected by means of software. The frequency of peaks and valleys provides respiration rate data, and intensities above a certain value are associated with cough reflexes. These processes involve taking a derivative of incoming data to detect maximums and minimums. This is easily accomplished by taking the differences between two successive digital measurements and comparing the value with previous and future difference values. The air volumes involved are determined by integrating the incoming digital data, which requires summation of the data over the respiration cycle.

For the method to be economical and statistically valid, it is usually necessary to monitor many fish on a round-robin basis. For example, 15 fish might be involved in the study and it might be desired to sample each fish for 1 minute, every 15 minutes. The sampling rate during one measuring cycle has to be 120 points per second. A real-time clock (one ticking independently of the computer) can interrupt the computer at line frequencies (60 or 120 hertz) and the computer can use this information to continuously update multiple registers so that it can keep track of time (see Fig. 4).

Analog-to-digital converters are available which can digitize the data in 10 to 100 μsec, and the data may then be input by means of digital input interface modules. Because of the mathematics and data storage required, it is desirable to

Fig. 4. An analog-to-digital and digital-to-analog conversion experiment.



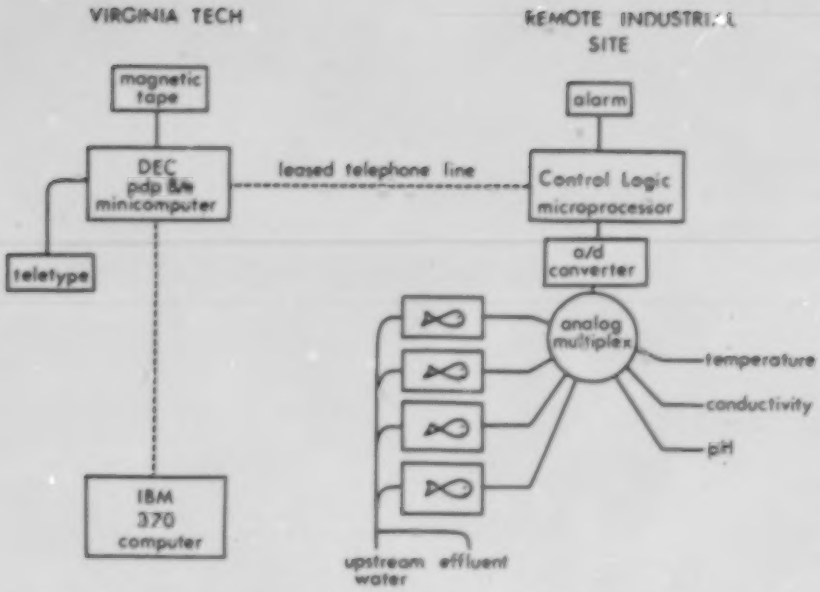


Fig. 5. A pollution monitoring system. [Courtesy of J. Cairns, Center for Environmental Studies, Virginia Polytechnic Institute and State University]

have a machine with multiple accumulators and the ability to store data easily in arrays (one array per fish) giving respiration and cough data as a function of time. Reports would consist of typed forms produced every hour showing the time, fish number, respiration rate, and cough rate for the four quarter-hour periods, and the average.

The Motorola M6800 (Fig. 6) is a microprocessor that has two accumulators, a stack pointer, and the necessary program counter and flags. In addition to the standard zero, negative, and carry flags, it has an overflow flag. Most important, it has an index register, which makes array handling simple. In handling an array it is necessary to be able to use the base address (starting address of the array) and a displacement (how far down the array the desired element is) to calculate the absolute address (exact location in memory) of the element needed. An index register loaded with the base address is used for this purpose as follows (small capital letters indicate instructions given by the operator).

FIRST: CALCULATE RELATIVE LOCATION ELEMENT (DISPLACEMENT)

THEN: FETCH FROM MEMORY BY CALCULATING [(DISPLACEMENT) + (VALUE IN INDEX REGISTER)].

The value in the index register is not changed, and the CPU has another register, invisible or transparent to the user, in which the summed absolute address is entered to perform the correct fetch. It works just like the pointer-to-memory in the Intel 8080; indeed, the normal mode of addressing on the 8080 is index addressing with zero displacement. The Motorola M6800 can accept displacements of up to 128 places. It can also handle program counter relative addressing in a very sophisticated manner in comparison to the Intel 8080.

Mathematics

Most people employ a computer to perform calculations rapidly and accurately. Yet how the computer handles a mathematical operation determines the time required to perform it. This is not understood by the average user. The fact that an 8-bit-wide data word will only accommodate a range of integral values from 0 to 255 has been mentioned. It is possible in 8-bit processors to handle numbers in register pairs, and to think of data words 16 bits wide, accommodating a span from 0 to 65,536 (64 kilobytes). However, since only the accumulator register can perform mathematics in the CPU's mentioned, adding two 16-bit numbers (double precision in this machine) requires considerable juggling of the 8-bit bytes comprising each 16-bit word. If the numbers are in memory, as is usually the case, the number of operations required make this an extremely slow process.

An alternate notation, which will increase the range of numbers that can be handled, is like the scientific or floating notation commonly used. A number can be stored in 2 bytes (16 bits) in the following manner.

$$\frac{\pm\ \text{exp}}{\text{-----}}\ \frac{\text{fraction}}{\text{----------}} = \text{number} =$$

$$\pm\ .nnnnnnnnnn \times 2^{\pm nnnnn}$$

The range of permissible numbers can extend from

$$.1777_8 \times 2^{37_8} (\sim 32{,}000)$$

to

$$.1000_8 \times 2^{-20_8} (\sim .00001)$$

a dynamic range of about 3×10^9.

The ability to handle larger ranges and fractional numbers has been achieved. The price has been paid in lower resolution—only 10 bits for the fractional representation in floating point (1/1,000) but 16 bits in the classical integer representation (1/64,000). To increase the precision, more bits are needed in the representation of the number. Typical compromises involve 8 bits of exponent and 16 bits of fraction, including a bit to represent sign (1 = −; 0 = +).

In an 8-bit machine the handling of floating numbers is time- and space-consuming. The program (software) necessary to support conversions from integer format to floating point format, to support the mathematical functions of addition, subtraction, multiplication, and division, or to support any extended functions (logarithm, sine, and so forth) is long. Typically, 1500 bytes for four-function arithmetic and 3000 bytes for simple extended functions are required.

The scientist should be aware that although he can integrate, take logarithms, and so forth directly, the computer must perform all transcendental functions by approximation methods. In calculators this is done in hardware, and the times involved are often hundreds of milliseconds. In computers with minimal configuration this must be done by software. Typical times for minicomputer systems are 1 to 10 msec. For microprocessors this time must be multiplied by a factor of 2 to 10. Both the execution time of a code and the cost of writing that code must be carefully evaluated.

It should be noted that the four processors discussed in this article, if purchased as operating or near-operating systems, would each cost $1600 ± 20 percent in a configuration with 8 kilobytes of memory, terminal interface (serial I/O), and parallel interface (digital I/O). Price considerations might encourage a user intending to build 500 units to employ a less expensive CPU, despite the extra software costs, since he can write off program development costs over the entire

production run. A different set of financial factors is involved when only one or two automated units are being considered. Anything that will reduce software preparation costs is vital. It is for this reason that a careful analysis of your problem is important in matching CPU specifications to your needs. With this in mind, a more complex problem is examined next.

Application 3

The typical amino acid analzyer is a system ideally suited for automation by microprocessors. In operation a number of valves must be controlled by the processor to change the eluting buffer in both concentration and composition. Usually absorbance at several wavelengths is used to detect the eluting peaks. As the data are received by the computer they must be processed to reject information below a selected threshold value. When a peak is detected, rather sophisticated mathematical operations must be applied in real time. A running weighted digital filter is applied to the data to reduce noise. First-derivative techniques are employed to detect maximums and widths at half-height. Second-derivative techniques are often used to detect and help characterize shoulders. Because of ammonia bleed it is often necessary to accommodate baselines that drift and to subtract out plateaus.

In addition, there is no point in automating unless the unit is programmed to identify materials by the time window in which they elute, and calculate from calibration factors the concentration of each amino acid. This requires the system to be capable of accepting a command to calibrate, followed by a list of amino acids in the standard and their known concentrations. The amino acids are entered in the order in which they will be eluted from the column. As the sample is analyzed, time windows are calculated and response factors determined from the known concentrations and observed absorptivities. These constants are stored as arrays for use when an unknown is processed.

This requires a processor with good memory addressing capabilities, multiple accumulators, and a 16-bit data word to accommodate the mathematical operations required. Arithmetic capabilities available in hardware will be helpful.

The IMP-16 series (Fig. 7) of the National Semiconductor Corp. has four accumulators. Two of these (AC2 and AC3) may be used as index registers, allowing "simultaneous" access to two arrays. The remaining accumulators can be used for the typical mathematical operations.

The IMP-16 is a microcoded CPU. This means that a special decoding element capable of handling very long "command" words is present. Being longer, these commands can elicit quite complex operations with a single instruction. It is possible with microcode to provide the IMP-16 with the capability of implementing special instructions to perform double precision (32 bits on this machine) integer arithmetic (addition, subtraction, multiplication, and division). The two 16-bit numbers involved are placed in the AC0 and AC1 registers, and the microcode executed. The results are left in AC0 and AC1. This speeds up the mathematical operations considerably in comparison to software integral arithmetic. Numbers to 4.3×10^9 are accommodated.

The IMP-16 has the ability to address memory relative to the program counter. This can be viewed simply as another operation in which a displacement with reference to a base address is used to calculate memory address, but where the program counter provides the base address. This is a logical extension of the immediate addressing scheme described for the Intel 8080, where only zero displacements were allowed. However, in the IMP-16 addresses ±128 bytes from the program counter can be accessed. All of these schemes are attempts to provide methods by which fewer bits and less time are needed to access memory in a program.

The IMP-16 also has indirect addressing. Instead of a direct, relative, or indexed operation giving the address in memory required (pointer-to-memory), the indirect operation gives the address of the address in memory required (pointer-to-a-pointer-to-memory). This feature is valuable because the stack in the IMP-16 is a hardware stack limited to 16 words. This makes access to materials on the stack rapid, compared to memory fetches, but one cannot build long arrays in the stack without data "falling through" (more than 16 words put in the stack) and being lost.

The indirect method allows a pointer to be set up in a memory location and then used to create an array in another part of memory. Unfortunately it is the program's task to continuously increment or decrement this pointer location as the file is created. This resembles the operations involved in the Intel 8080, where it was the user's job to update the pointer-to-memory register in creating a list.

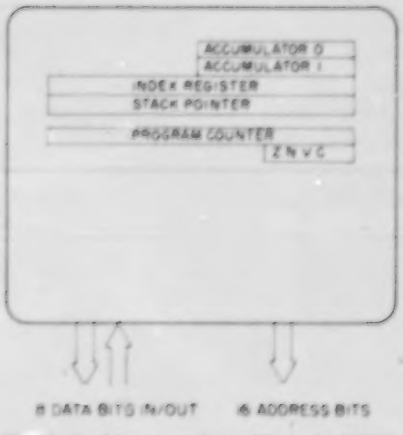


Fig. 6. Architecture of the Motorola M6800 CPU.

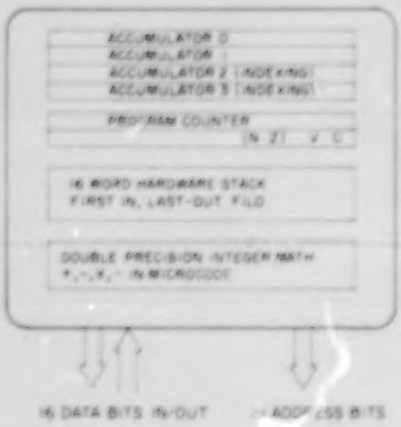


Fig. 7. Architecture of the National Semiconductor IMP-16 CPU.

Application 4

Since the lifetime of perfused biological samples is short, it is usually necessary to abstract as much information as possible from the experiment in a short time. Typical experiments involve nerve or muscle preparations in which a variety of transducers are placed in a large number of places (Fig. 4). Galvanic activation of the preparation, controlled in intensity by the computer, can lead to a response studied by simultaneous measurement of the rate of propagation of the electrical impulse in the nerve and the response at the activated tissue.

Electrical, specific chemical ion, and pressure transducers will be needed to obtain data. These data can measure (i) rates of electrical propagation of the stimuli, (ii) propagation delay areas, (iii) ion

Table 2. Memory and time needed to add two 16-bit numbers.

Microprocessor	Memory (bits)	Time (μsec)
Intel 8080	120	46
Motorola M6800	104	35
National Semiconductor IMP-16	80	28
Digital Equipment Corp. LSI-11	48	11.7

recovery rates, and (iv) response time in the tissue. Predictions of fatiguing rate must be made, and closed-loop control of the experiment may be desirable.

Sampling of data in such a system requires a processor with extremely well developed memory addressing, since multiple arrays will be created and used. Mathematical operations are quite sophisticated and require real-time execution with a fair degree of resolution. It would probably be desirable to be able to replay an experiment in slow motion on a video terminal. Heretofore these operations have been the perogative of a large minicomputer, but microprocessors are now available with the capabilities of handling the problem.

The LSI-11 (Fig. 8), manufactured by Digital Equipment Corp., is a microprocessor chip set that emulates the operations (instruction set) of Digital Equipment's PDP 11/40 computer. It is thus a CPU that falls in the gray area between microcomputers and minicomputers. With six accumulators, any and all of which can be used as index registers or stack pointers, the LSI-11 has remarkable power. It has all the forms of addressing previously described: direct, indirect, index, and relative. In addition, it has autoincrement and autodecrement features when indirect addressing is used. In this mode the CPU automatically steps the pointer forward (increment) or backward (decrement) through a sequential operation. This removes the need for the program to update pointer-to-memory information as contiguous files are being created. It also has microcoded integer multiplication and division as well as two-word floating arithmetic addition, subtractions, multiplication, and division. The latter mode will accommodate a range of $\pm 10^{\pm 38}$. Typical time for floating division is 150 microseconds.

Discussion

It is appropriate at this point to sum up the capabilities of the various processors by comparing the relative amount of memory storage (for program) and time that would be needed to execute the addition of two 16-bit words (see Table 2).

For 32-bit integer multiplication or division the IMP-16 will take 2.7 times more time than the LSI-11. Although equivalent benchmarks are hard to find, a 3-byte software floating addition takes 2.5 msec on the 8080, compared to 50 μsec for a 4-byte firmware floating addition on the LSI-11—a factor of 50. Before attempting to use this information alone to focus on the processor of choice, let us look at software.

The generation of software for microprocessors is, and will continue to be, the major problem. The CPU can accept code only in a language made up of 1's and 0's—machine code. Humans "speak" arithmetic, algebra, and words, but not 1's and 0's. It is possible to "hand code" from an instruction card in machine code. In the simpler processors like the Intel 8080, it is possible to code programs of less than 500 bytes in this way—if the user has had considerable experience. As the programs become longer, mistakes in address calculation and the need to insert forgotten code become insufferable problems. In more complex machines, particularly where displacement values must be calculated, the frustration point comes even earlier.

It is possible to operate in a mnemonic language, with abbreviated symbols to assist humans to remember instruction codes. In this method the user writes source code (a language from which another language is to be derived), using an editor which allows for corrections, insertions, and deletions. This source code can be stored temporarily on paper tape, cassette, or high-speed rotating memory such as floppy disks. The latter is the only really satisfactory solution because of problems of handling, speed, and error rate in the other devices. The source code is then submitted to another program called an assembler, which does all the address calculations and converts the words into instructions that can be handled by the machine. Unfortunately, adding the software to support the editor, assembler, and disk storage and retrieval to the bulk storage hardware increases the cost of a microprocessor to a point where it is equal to or greater than that of some minicomputers. Where large numbers of units will be manufactured in-house or for resale this can still be economical, but for the one-of-a-kind developer it is a disaster. It would be better to start off with a well-known minicomputer which has a large amount of available software.

Most laboratories are using cross-assemblers (assemblers running on one machine that produce machine code for another) which run on larger minicomputers or through time-sharing facilities. These are available for most of the large-volume microprocessors; however, a careful study should be made before assuming that a cross-assembler is indeed available for your processor, and that it is "bug-free."

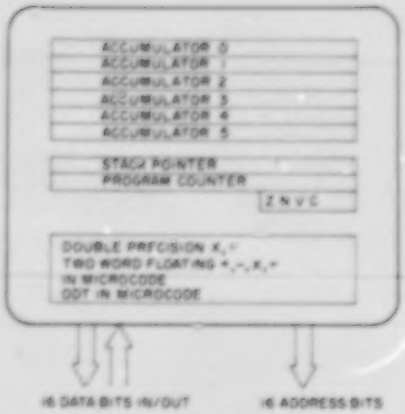


Fig. 8. Architecture of the Digital Equipment Corp. LSI-11 CPU.

Finally, several companies are beginning to offer higher-level languages such as Basic or a subset of PL-1. The ability to handle real-time functions is lacking in the former (just as real-time Basic lacks many I/O features). Both produce code that is inefficient in its use of memory and therefore costs time and money to execute. This will not be a serious factor for the one-of-a-kind developer. He should not be stressing the CPU, and a few hundred dollars in memory will easily offset the price of coding, which costs about $10 per line. At this point in microprocessor development high-level languages are not universally available or properly developed, and they may not be available for the CPU that meets the hardware and software characteristics needed for your application. At least three vendors are making microprocessors that use the same machine code as an existing minicomputer. Thus, software at any level can be developed on a host minicomputer and "downloaded" without code conversion. This is one elegant feature of the LSI-11 which is a mini-based microprocessor.

It is vital that the prospective microprocessor user be aware that hardware and software documentation will be far less extensive than similar material provided with minicomputers. Microprocessor vendors assume a certain level of sophistication in their end users and provide accordingly. Microprocessor

hardware may be purchased in the form of CPU chip sets, but this is not recommended for the average scientist. Nor is purchasing the CPU card that contains most of the decoding and I/O control elements. A complete system with some software support is most practical. For the novice, or even those with considerable experience with high-level languages, it is essential that the machine have a bootstrap (the program that loads subsequent programs) and a rather sophisticated debug program. The latter permits the user to examine and alter the content of memory addresses in machine code. Programs can be initiated and in some cases run in single-step mode or with set breakpoints. These are the only feasible ways to find errors. Diagnostic software that detects and reports hardware malfunctions is essential, since few microprocessors sold have field or factory repair service.

Although high-level executives and operating systems (a collection of programs that aid in programming) are advertised, the state of the art is still primitive. Since much of the information published in this area for minicomputers is often misleading, it is doubly important to establish the reliability of software claims in the microprocessor area.

The best way to judge your chances for success in the plunge into microprocessors is to see if the prospective vendor or you can understand and answer the following questions about the system of interest to you.

1) Can the firmware bootstrap and debugging program accept terminals with mark, space, odd, or even parity?

2) Does the debugging aid permit the saving of all registers and flags during execution of breakpoints?

3) Are the binary-formatted tapes provided in absolute format, or is a linking loader needed? If one is provided, how much memory is required to run it?

4) Is there a software interrupt disable command? If the interrupt priority scheduling is in hardware, does it arbitrate correctly? If it is in software, does masking-off capability exist in the interface hardware?

5) What is the resolution of the floating point package? (Some 24-bit floating point routines only give five-digit resolution, compared to eight digits on handheld calculators.)

6) If an operating system is supported, are the device handlers written so that they may be added and deleted easily to reconfigure the system—or are they written in in-line code, making alteration almost impossible?

7) If high-level languages are supported, how long a user program can be run, and how many user-assigned symbols are permitted in the normal memory configuration?

8) What is the bit drop-rate of the peripherals provided?

9) Does the vendor provide ROM "burning" facilities so that programs may be loaded into nonvolatile memory for long-term use?

10) Who is responsible for issuing refresh commands in memories that are dynamic? And does this process interfere with the operation of other peripherals?

All of these above questions have been constructed from actual case histories in microprocessor applications. The problems involved were overcome, but at considerable cost in time and money. If the vendor or you cannot respond to these questions, you are not ready to use microprocessors. Microprocessors offer an inexpensive solution to the hardware problems encountered in laboratories, but they may engender a manpower expenditure in programming and hardware setup that is prohibitive.

Developments in this area have been breathtaking in the last 3 years, but a technological plateau has been reached that I believe will be maintained for several years. The main thrust during this period will be toward lower cost rather than an increase in power that is evident to the end user. One can then expect the introduction of new capabilities in hardware, mathematics and statistics packages, memory methodology, and substitutes for rotating bulk-storage devices. Large-scale integrated circuit chips will ease the interfacing task by providing flexible, inexpensive I/O control, but they will usually be specific to particular CPU's and peripherals. A supermarket or erector set philosophy in microcomputer system design is still a long way off. Some think it will never come.

By this point, it is hoped that the reader will have enough understanding to analyze the potential application of microprocessors to his tasks. In processes where mathematics and memory accessing are minimal and the programs are of a length suitable for hand-coding, the Intel 8080, with its simple-to-understand instruction set, is ideal. Where speed and language sophistication are important, CPU's like the Motorola M6800 are applicable. Where speed, mathematical power, and instruction set versatility are important and expansion is envisaged, look to mini-based microcomputers like the LSI-11.

It is strongly urged that your first introduction to computers not be microprocessors. The learning curve is a steep one that is best approached by some experience on gentler slopes and some previous exposure to such factors as the cost of program development, problems with the mathematics package, and running out of real time is desirable.

Notes

1. Suggested further reading: A. Osborne, *An Introduction to Microcomputers* (Adam Osborne and Assoc., Inc., Berkeley, Calif., 1975); R. H. Cushman, "Second annual microprocessor directory," *EDN* **20**, 42 (20 November 1975); R. E. Dewy, P. J. Van Vuuren, J. A. Titus, *Anal. Chem.* **46**, 917A (1974); *ibid.*, p. 1055A.
2. This article has been developed with the aid of E. Fiorino, J. Fiorino, P. Hallman, J. Leone, J. Starling, D. Binkley, D. Hoxley, C. Knipe, W. Nunn, D. Shearer, M. Starling, and H. Wohltjen. These people compose the Chemistry Department's Instrument Design and Automation Research Group.

149

Microprocessor Application: A Less Sophisticated Approach

James T. Arnold

There is a widespread appreciation for the impact of data processing by digital computers, which profoundly touches so many aspects of life today. The rapid development and application of the digital computer has indeed been one of the most dramatic achievements of technology. Digital methods of management and manipulation of information at very high speed with digital computer operating systems give access to solutions of problems so vast and so intricate as to have been simply inconceivable before their advent.

The unique capability of the digital computer operating system to solve complex problems in a timely way justifies the necessarily enormous expenditures for its development and its services. These high costs have largely restricted the application of computer methods to tasks which could justify the fees by efficient use of the full capability of the computer operating system.

The microprocessor offers the means to relieve this cost constraint in the use of digital methods for problems of smaller scope. It is a revolutionary tool by virtue of the fact that it incorporates the potential performance of at least the lower end of the minicomputer line and yet, as a component, it has a cost so low (*1*) that its use can be justified in applications requiring only a small fraction of its capability. Moreover, while the application of complete computer systems, not to mention their design and construction, has been largely and properly left in the hands of a professional cadre of specialists, the microprocessor, intelligently used, can be an effective tool in the hands of the reasonably talented amateur.

This suggestion is not to denigrate the need for professional attentions in major computer problems. (The microprocessor has grown out of computer technology, and it reflects the complexity and sophistication of that field.) It is rather to state that the microprocessor offers, incrementally and at low cost, digital management and manipulative capability which can be understood and implemented, for simple problems, by nonspecialists. Applications in which use of the full range of even the minicomputer is contemplated are probably still served best by the engagement of professional technologists. The rare amateur who succeeds in such an application will, himself, have become an expert, at the expense of a considerable effort.

In this article I view uses of microprocessors as elements of instrument systems or to facilitate individual experiments. Of course, the instrument or experiment contemplated must succeed on its own merits; designs including microprocessors do not automatically guarantee superior performance. Indeed, because of the programmed nature of microprocessor operation, the experiment or instrument designer must exercise considerable care in planning and specification to avoid constraints that would be inconvenient in later operation.

Another Description of the Microprocessor and Its Operation

All descriptions of the microprocessor inevitably tend to illuminate its role as a potential element of a computer system. The computer role is deliberately deemphasized here in favor of such descriptors as controller and data manager.

The microprocessor is a digital electronic device of considerable complexity in which electrical current pathways are suceptible to controlled switching. It is designed to operate cyclically in such a way that during a cycle, data are (electrically) abstracted from one or more sources, manipulated by an arithmetic logic unit, and then distributed to one or more destinations, these steps of the cycle being under control of a prescribed control regimen called an instruction (Fig. 1). The data are represented in this process as ordered sequences of binary statements. The binary nature of each statement, commonly called a bit, is conventionally conceptualized by word pairs such as false, true; off, on; or zero, one; and the ordered sequences are connoted as words or as (binary) numbers. The data themselves can be as diverse as definition allows.

Implicit in the utility of the data is some means for their temporary or permanent storage. Storage can be accommodated in registers, either external or internal to the microprocessor. A single register is a component possessing an ordered number of cells, each of which is capable of reflecting a binary statement in such a way that the statement can be entered or abstracted at will. The order of the cells in the register and the identity, or address, of the register itself preserve the identity of the data.

The microprocessor system, either internally or externally, is provided with several special purpose registers in addition to general purpose registers for data storage. The control regimen or instruction for each cycle of operation is impressed on a first special register, called the instruction register, in the form of an ordered sequence of binary statements—that is, an instruction word. A second special register is called the program register. This register contains a binary number which identifies the location from which an instruction, or occasionally other information, is to be drawn.

In addition to executing a cycle prescribed by the instruction word, the microprocessor system facilitates the execution of a sequence of instructions by adjusting the content of its program register, thereby calling for another instruction. Figure 2 illustrates this operation, showing three successive cycles. The set of instructions, called the program, shown in the left-hand column of Fig. 2, actually governs the events taking place in the microprocessor. These events, shown in the center column, include the step of adjusting the program register which points to the address of the next instruction. The complete collection of instructions for operating the microprocessor is called the program. The microprocessor cycles shown are numbered in the sequence P, $P + 1$, $P + 2$. It is quite natural to assign addresses in the program as sequential numbers as well, and to interpret the content of the program register as a multiple-digit binary number. The most usual adjustment of the program register then, is to in-

The author is a senior scientist in the Corporate Research Laboratories, Varian Associates, Palo Alto, California 94303.

crement this number by one. (It should be noted that an instruction may direct the microprocessor to treat the next program address as the location of a data source rather than an instruction, and another may direct the entry of a nonserial address pointer in the program register.)

In summary, the microprocessor is an example of a stored-program device; it can proceed autonomously through a sequence of operations, selecting, manipulating, and distributing data in accordance with a previously stored plan or program. It can perform its function rapidly; cycle times are typically from 1 to 20 μsec.

Evolution of a Simple System Using a Microprocessor Well Below Its Capacity

The following paragraphs describe the incorporation of a variety of functions into an imaginary instrument by the use of a microprocessor. Actual development of a system provided with such a set of functions should be undertaken as a whole rather than by the steps that are serialized here for illustrative purposes.

Keyboard and display. The imaginary instrument will involve interaction with an operator; it will require inputs from the operator, and it will display certain responses to these inputs and its own operation. The growing popularity of keyboard entry and alphanumeric display suggests selection of this method to interface the instrument to the operator. The first function of the microprocessor, under control of a stored program, is to accept keyboard entries and to output these, and possibly other information, to a display.

The operation of a microprocessor lends itself naturally to this function. The operator's keyboard can be treated as a data source and the display as a destination. Successive access to source and destination requires control by a repeated sequence of program instruction steps. Particular requirements of the display device may necessitate the use of manipulative as well as distributive instructions.

Implementation of the keyboard-to-display function is shown diagrammatically in Fig. 3. A word of explanation is required for the blocks in Fig. 3 marked "memory." Memories are generally external to the microprocessors and consist of large numbers of registers or other word storage devices, each having a unique address susceptible to call from the microprocessor's program register output. The read only memory, ROM,

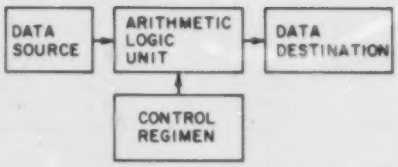


Fig. 1. Schematic representation of a data flow cycle in a microprocessor.

acts exclusively as a source of instructions and data, being loaded with the proper sequence of words. The random access memory, RAM, can act both as a data destination and as a source to recall data previously deposited.

Operation begins with an instruction in the ROM which directs the processor to

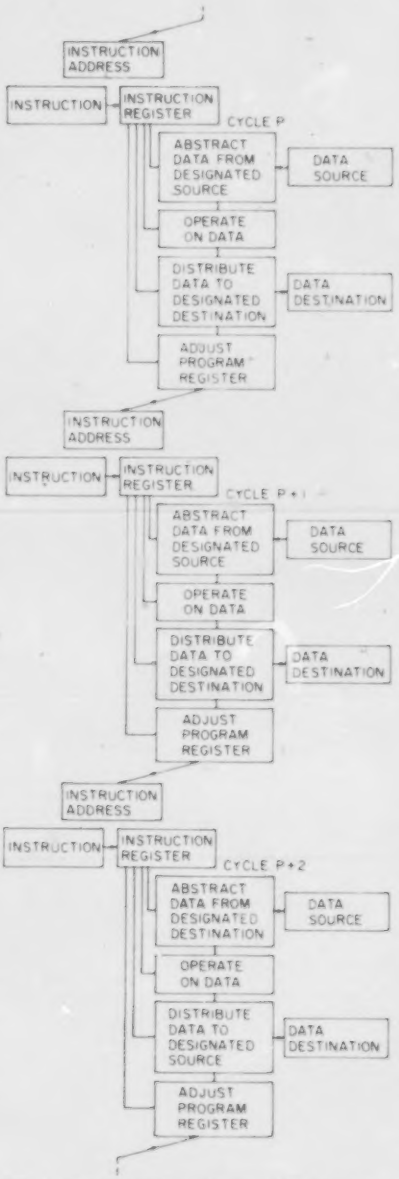


Fig. 2. Representative progression of microprocessor cycles.

"read" the first entry from the keyboard. When the entry is received, the microprocessor must perform the equivalent of the following steps.

1) Retain the entry in temporary storage, for example, the RAM.

2) Assign a location for the entry on the display.

3) Retain this assignment paired with the entry itself in temporary storage.

4) Manipulate the entry to conform with a code appropriate to the display.

5) Deposit the coded entry in the assigned display location.

Many displays are volatile, and in this case the last operation must be repeated at a rate compatible with visual persistence (about 30 sec^{-1}). As the successive keyboard entries are directed to the display, each must be updated in its turn at this rate, and to accommodate this requirement the microprocessor must rapidly generate an elaborate sequence of data to deliver to the display.

Fortunately the system can be organized to do just this. One simple method is to have the microprocessor arrange a data matrix in a block of the RAM such that its addresses are in a one-to-one correspondence with locations on the display. This display "mirror" is built up relatively slowly by entries from the operator's keyboard, as required, and is rapidly scanned serially by the microprocessor to provide the continuous, complex (but almost mindlessly repetitive) data stream required by the display.

A cathode-ray tube with a conventional television raster is a popular display device. The data stream, encoded by the microprocessor according to a stored program in the ROM, can be made to accommodate the image dissection and timing requirements of the raster. Factory encoded character generating ROM's are even available, at low cost, for this purpose, relieving the instrument designer of the chore of generating the codes.

While the function described above does not significantly tax the operating sophistication of the microprocessor, and the rate of data entry from the keyboard is extremely slow by microprocessor standards, the continuous update of the display screen will burden the microprocessor to the point that no other functions can be accommodated. This difficulty can be relieved by augmenting the system of Fig. 3 with special purpose hardware which, standing alone, can convert the contents of the RAM display data matrix block into the data stream required by the display. This augmentation is shown schematically in Fig. 4. Access to the shared display RAM is invoked occasionally for data entry by

the microprocessor and continuously by the dissected image generator for display updating. To prevent ambiguity, access to the shared RAM address is normally continuous from the dissected image generator, but is preempted by the microprocessor whenever the latter is required to enter data. This arrangement for priority of access permits completely asynchronous operation of microprocessor and display. With the hardware augmentation in the system of Fig. 4, the microprocessor is grossly underemployed, and other functions can be accommodated.

Instrument control and data acquisition. Since the function of writing to a display, by itself, is not an adequate operator-instrument interface, the excess capacity of the microprocessor can be turned to good use in performing the additional tasks of instrument control and data acquisition. Instrument operating programs and operating parameters can be stored in the ROM, and instrument output, converted to digital words, can be read and stored in an orderly way in the RAM. Use of the keyboard and display can be merged with these tasks. With the keyboard, the operator can access the microprocessor to select operating programs and parameters for a desired instrument cycle. Under control of the selected operating program, the microprocessor can also be directed to accept progressive and final results from the instrument, to format these results, and to enter selected results on the display.

An implementation of this more complicated operation is shown schematically in Fig. 5. Data paths to and from the instrument and an additional RAM have been added, and the ROM must be considerably expanded to accommodate the new tasks required of the system. The utility of the system is dependent on the care and imagination exercised in the hardware and operating program design, and the burdens and benefits of microprocessor use are most sharply focused in the design process.

In terms of their hardware needs, instruments require input intermittently or continuously, and they generate outputs which must be accepted to formulate results. The inputs and outputs may cover a variety of physical quantities, not necessarily electrical in expression, and generally analog in nature. Since the microprocessor is digital and electrical, transducers are generally required to convert data to appropriate form in either direction. The transducers adjacent to the instrument are likely to be analog in character and may relate electrical levels to other physical quantities. The ones adjacent to the microprocessor accommodate analog-to-digital and digital-to-analog conversion (ADC and DAC).

Devices to accomplish a wide variety of conversions are readily available on the commercial market, specified with reasonable intelligibility. In digital parlance, the precision of a device is usually specified in bits, each bit representing a power of 2. Thus 8-bit precision represents a resolution of 1 in 2^8 or 1 in 256. Other specifications include speed of conversion, absolute accuracy, range, and so forth.

Data transfers are negotiated one word at a time. The width of the word (in bits) is determined by the data paths; the data paths into and out of most microprocessors are 8-bits wide, although other widths may exist in other parts of the system. (This constraint does not limit the precision of the data, since more than one word may be used to transfer a particular element of information if greater precision is required.)

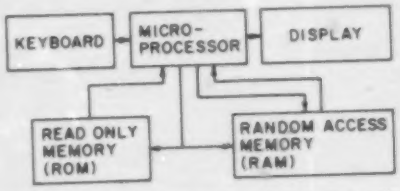


Fig. 3. Schematic implementation of the direct keyboard-to-display function with a microprocessor.

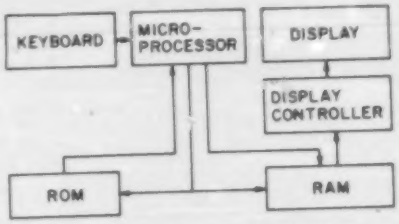


Fig. 4. Addition of an auxiliary display controller to relieve the microprocessor of repetitive display refreshment.

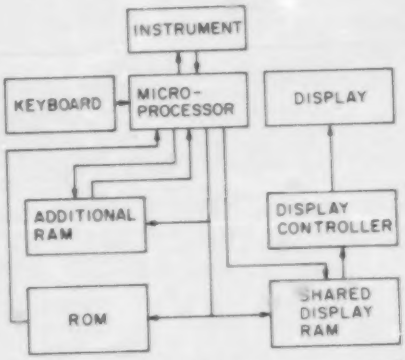


Fig. 5. Use of the microprocessor to control an instrument and accept resulting data directly.

Each of the functions to be controlled and each of the quantities to be monitored by the microprocessor must be assigned an individual address. Most microprocessors are provided with input and output procedures which represent the means to communicate with external devices. In addition, the addressing structure, usually designated for memory exchanges, can also be used to access a number of external registers for data communication to and from an instrument.

Equal in importance to the hardware in the use of a microprocessor in an instrument system is the organization of the operating programs. The designer must devise through these programs the means to utilize as much of the capability of the instrument as possible, to improve the efficiency of its use, and to reduce the errors and ambiguity of its results. To do this he must have a thorough understanding of the operating characteristics of the instrument and the methodology of the measurements made. Once a set of operating procedures has been established for a desired operating program, the procedures can be reduced to a chain of operating steps, each of which can be negotiated by the microprocessor under control of one or more instruction words.

The chain of operating steps can be as elaborate as the instrument requirements dictate. It can branch to alternative lines on the basis of incremental results. It can be keyed to external parameters such as real time. It can be interrupted temporarily, on demand, for the service of off-line requirements. It can sense anomalous events and return control to the operator, signaling a diagnosis of the anomaly it has acquired. The orchestration of this operation is by means of the flow of instruction words called from the stored operating program by the microprocessor. Although the reduction of the operating procedures to a set of instructions is necessarily complicated, it is negotiable, by hand, for simple problems, and writing the program can be facilitated by design and organizational aids such as flow diagrams which map the sequence of program steps.

The designer may choose, in simple cases, to construct the program in the instruction code directly understood by the microprocessor. In routines of greater extent, coding aids are available which can be accessed by an accumulation of program statements in a simple formalism called assembly language. The

reduction of these program statements to microprocessor code requires the use of a program assembler routine residing in a modest computer operating system. Programming aids of even greater abstraction are available; however, their use entails access to specialists and large computer operating systems.

Most instrument operating routines are relatively simple and can be approached even at the level of microprocessor instruction codes. An example may be found in the frequently encountered case of observing a dependent variable quantity as a function of the sweep of a predetermined small interval of an independent variable, making observations at uniformly incrementing values of the independent variable. The procedure is as follows.

1) Begin, set initial and terminal values for independent variable in temporary registers.

2) Set memory (storage) pointer to the head of a one-column data matrix in the RAM.

3) Acquire the value for the dependent variable.

4) Deposit the value in the data matrix (RAM) address.

5) Increment data matrix address.

6) Increment independent variable.

7) Compare independent variable with its terminal value.

8) Exit if done; else

9) Loop back to step 3.

This small procedure is generally embedded in a larger routine which addresses, for example, the same measurement for a different interval of the independent variable. The larger routine itself may be embedded in a still larger program which (by keyboard selection) can access this or other routines. As long as the routines are simple, they may be individually coded with relative ease, and the collection of routines can be made accessible on selection by a supervisory portion of the program.

There is no rule other than good sense to proscribe hand coding as the problems grow larger. At some point—sometimes identified, in the end, by sheer frustration—the designer will discover that hand coding is too costly in time and effort. At that point a more formal approach, involving abstract methods of assembling microprocessor instructions with a computer operating system, will be clearly indicated.

Specialized instrument controllers. The simple instrument operating routine whose nine steps were sketched above can be viewed as an advantageous use of a microprocessor in many cases. However, where the routine must be continuously exercised to operate the instrument, it may begin to strain the throughput capability of the microprocessor. In the interest of underutilization, this difficulty can be relieved by the interposition of an instrument controller, which accepts broader and less frequent control commands from the microprocessor and which accumulates results in such a fashion as to allow delivery to the microprocessor with the minimum of transmission consistent with integrity of the data.

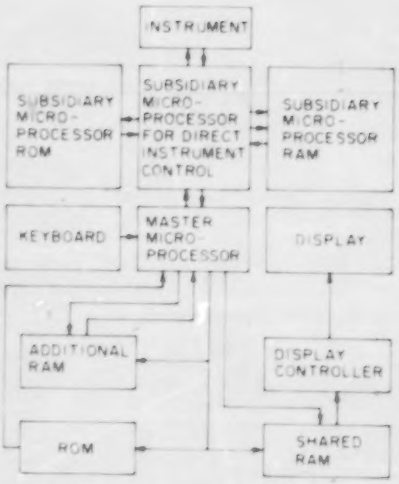


Fig. 6. Complete system for microprocessor instrument operation including the use of a subsidiary microprocessor, having its own ROM and RAM, for direct instrument control and data reception.

Such a system is shown in Fig. 6. In accordance with the example mentioned, the iterative part of the (independent variable) control routine can be managed by the controller with the resultant (dependent variable) data being accumulated temporarily within the controller itself. Transmission of the essentials of these data can then be called for by the microprocessor after each complete cycle of the routine. With such a controller, the microprocessor is relieved of continuous operation of the instrument and is required only to set the parameters defining the fiducial parts of the routine and to collect the relevant results.

The controller can frequently be so simple as to be amenable to design with hardwired digital logic, or even with analog circuits and switches. However, the reader will recognize that the controller, with no other obligations than to carry out a repetitive cycle of operation, may itself qualify for implementation with an underutilized microprocessor having its own ROM and RAM. This subsidiary microprocessor can then be operated as an instrument controller under the supervision of the primary microprocessor. Interconnection of the two through standard input/output data paths need not pose a difficult problem beyond the avoidance of possible timing conflicts which might lead to garbled results.

Data management and processing. With the system configured according to Fig. 6, where the primary microprocessor is underutilized in its function to respond to keyboard commands, to operate the instrument, to accept instrumental results, and to load the display, there will be capacity remaining to perform the function of more elaborate management and manipulation of data. The first task for the microprocessor in this area is perhaps simply one of arranging the data in a convenient format. Other tasks depend, of course, on the requirements of the instrument, and the operations described below are mentioned only by way of example.

The microprocessor can easily be programmed to sort data by a prescribed rule; for example, to segregate maximum and minimum values. It can also generate difference tables and locate turning points. The latter operations are arithmetic and invoke the computational capability of the microprocessor. Other simple arithmetic operations might include generation of moving averages or simple digital filtering of data.

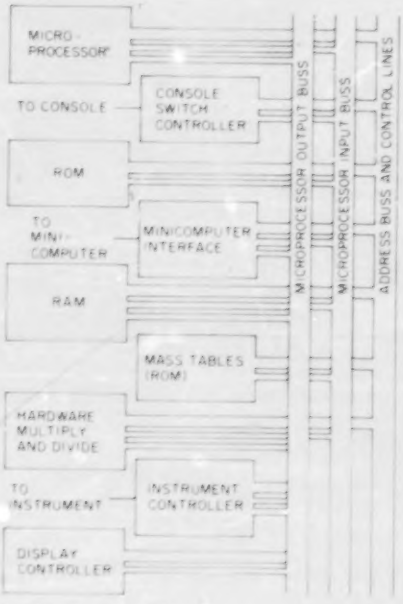


Fig. 7. System designed to operate a portable mass spectrometer with a microprocessor. The minicomputer interface allows the system to communicate with a minicomputer for checkout and program composition.

Even more complex data processing is, of course, possible. However, elaborate data processing is difficult to program without the help of programming aids, and it is advisable to enlist the help of specialists and to review the suitability of the microprocessor (compared to a minicomputer operating system) where the application involves elaborate computations. The use of a microprocessor in complete imitation of a minicomputer operating system with major data handling peripherals is likely to be more costly and will certainly involve more labor than the use of its model. Notwithstanding this precaution, the microprocessor can be employed to remarkable advantage in systems which, superficially, appear to be very complex.

Judicious distribution of hardware assistance to the microprocessor can reduce its load, as has been shown by the examples above. In the computational area, two commonly required functions which, if relegated entirely to the microprocessor, can burden it excessively and consume inordinate amounts of time are multiplication and division. A hardware auxiliary for fast multiplication and division can be devised, either along classical calculator lines or with special purpose large-scale integration digital components. To use this auxiliary, the microprocessor simply distributes the multiplier (divisor) and multiplicand (dividend) to the auxiliary device, and then reads the product (quotient) resulting. The device can be made to operate at such a speed that there need be no hesitation in the entire process.

In the programming area, algorithms have been developed in the technology of computer application which can be borrowed for microprocessor use to speed up and simplify other operations. These are too numerous to permit a comprehensive list here; however they include such functions as extracting roots, double precision operations, floating decimal point management, arithmetic and trigonometric conversions, and exponentiation. In addition, where special functions or calibration corrections are required, they may be accessed by simple interpolation of tabular entries introduced in blocks of the ROM or RAM.

Concluding Remarks

Since the microprocessor has been designed, by and large, to provide the principal part of the central processing unit of a computer, there has been a tendency on the part of vendors and many users to view it exclusively in that function. Good computers have been and will be constructed with microprocessors. In the paragraphs above I have tried to convey an alternative view, which perceives the microprocessor as a component having many features that can be used individually, and to a considerably lower level than its capacity would allow, without requiring the expertise of the computer architect. The very low cost of today's microprocessors and their ability to perform a variety of operations on digital data in accordance with a preconceived program can justify underutilization of their fullest capability.

No specific mention has been made of circuit details or of detailed instruction formats or algorithms in programming. Detailed data sheets are available for most of the processors offered (*2*). These give sufficient descriptions of microprocessor operating cycles, instruction sets, and other features to allow the enterprising amateur to develop useful simple systems. Successful use of the devices will often involve their augmentation in the systems with specialized hardware auxiliaries. These auxiliaries serve to reduce the complexity of the tasks required of the microprocessor itself.

A final illustration, Fig. 7, shows the advantageous application of a microprocessor to a portable mass spectrometer (*3*) which was designed and constructed by an amateur oriented to the approach suggested in this article. The microprocessor used was a first-generation device with a relatively slow duty cycle, and the system could not have been operated without the several auxiliaries shown. The system accommodated several modes of operation based principally on the monitoring of a number of selected mass peaks and displaying a processed evolution of their amplitudes in the form of moving average values. Included in the program were an automatic self-calibration routine and one to allow sequential scans of the entire mass range of the instrument. The stored program occupied 1792 words of ROM, and 1024 words of RAM were provided, primarily to accommodate variable program and a data matrix for raw and processed data.

Most of the reference material for microprocessors as devices is in the form of product release descriptions and manufacturers' data sheets. A useful summary list of current offerings of manufacturers may be found in one of the electronic trade journals (*4*).

The decision to undertake the design and construction of a nontrivial system using a microprocessor is a commitment to a major effort and should be contemplated with a good deal of circumspection. However, if the system is reasonably confined in scope, the essay can be rewarding and fun.

References and Notes

1. Texas Instruments Inc., Houston, recently announced a price of $21.15 each in lots of 100 for the 8080 microprocessor. This 8-bit device is very competitive in specification and has excellent support in software.
2. For example, Intel Corp., Santa Clara, California: *Data Catalog 1975*; *MCS-8 Microcomputer Set Users Manual* (November 1972); *MCS-80 Microcomputer Systems Users Manual* (July 1975). Motorola Semiconductor Products Inc., Phoenix, Arizona: *M6800 Systems Reference and Data Sheets* (May 1975). National Semiconductor Corp., Santa Clara, California: *Pace Technical Description, 1975*. Rockwell International Corp., Anaheim, California: *Introduction and Description, Parallel Processing System (PPS8)* (Document No. 20164N40, 1974, 1975).
3. J. T. Arnold and P. P. Robbiano, *U.S. Army Tech. Rep. No. LWL-CR-03P68C* (June 1974).
4. R. H. Cushman, "Second annual microprocessor directory," *EDN* **20**, 42 (20 November 1975).

Humanizing Computerized Information Systems

Guidelines developed in a series of workshops are presented and their implications are discussed.

Theodore D. Sterling

The accumulation and control of information is a critical function for government and private, industrial and nonindustrial organizations. Yet the role of information as an organizational resource is not very well understood, especially as it is related to the organization's environment. What does appear is that computerized information systems have become a facilitating technology that interacts with organizational, historical, and environmental pressures and goals to shape not only the internal structure of an organization but also its interactions with society (*1, 2*). There is little doubt that the computerized or automated information system is revolutionizing the management of most, if not all, systems by which goods and services are produced or information is accumulated. This shou'd be a source of great concern.

Weizenbaum (*3*) asked whether large computerized systems can be used by anybody except governments and really large corporations and whether such organizations will not use them mainly for antihuman purposes. The power of computerized information systems to control large enterprises answers the need to manage large systems and make them amenable to human control. By any criteria of management performance, computerization of a system permits its detailed control, and thus the computer is the ideal management tool. But the cost of the control is high. Start-up costs to redesign and computerize large-scale enterprises are immense. In concentrating on feasibility and workability and simultaneously minimizing costs, few systems designers seem to have been concerned about whether their products will be used for antihuman purposes.

In many ways, it is immaterial whether control over the management network is exercised by manual means or by automation. As long as official procedures are detrimental to human dignity, nothing is changed in converting to automation—except that individuals may shift the blame for their oppression from the human cog to the computer cog. It may be necessary, therefore, to clarify the dehumanizing components of a management system, which may be present whether or not the system has been automated, and to provide relief for any suffering they may have caused.

In a previous analysis (*4*) I pointed to two design strategies that account in large part for the presence of dehumanizing features in a management system. First, the efficiency of an enterprise is commonly increased by treating the recipients of the service and participants in the system as unpaid components whose time, effort, and intelligence do not appear in the cost accounting. Then, in order to maintain the efficiency of procedures once they have been established, the system is made exceedingly rigid, permitting freedom of action at only a few, usually hidden, focal points of real control. Dehumanizing features are thus already ingrained in most systems of management, and automation of such systems simply transfers the dehumanizing practice from one means of exercising control to another, codifies it in computer programs, and expands its influence to a larger circle of recipients and participants. To provide for the smooth and efficient operation of a largely computerized management system, the automation process makes demands of its own on all participants which decrease the area of free action remaining to the individual. Rules of procedure are thus dictated by the growth of machines and not by the needs of man. As a consequence, it is possible for the machine to capture the prerogative to formulate questions important to man. If we take such developments as inevitable we are surrendering our humanity.

The point is that an intelligent understanding of a machine mode of control may be delayed until long after this control has been exercised. Wiener (*5*) argued that although procedures laid down to satisfy a process of automation are subject to human criticism and modification, such criticism may be ineffective because it may not surface until long after it is relevant. It may be too late then to correct the damage to the human condition. Systems are not detached from the people they interact with and the settings they create, and people strive for a sense of dignity, have needs that should be taken seriously, like to be treated with consideration and courtesy, and occasionally act as individuals—in short, they are entitled to be treated as human beings.

Despite the overriding importance of a person's dignity and humanity, little is known in terms of "scientific" specifics about the operational meaning of these concepts or the antecedent conditions that enhance or diminish them. Relatively few analyses have been devoted to systems features that may humanize organizations (*6, 7*). We know of only one attempt to incorporate humanizing features in a system and to evaluate their effects (*8*). Yet we cannot afford to wait for knowledge to accumulate about the procedures to be incorporated in information systems or information parts of systems to help avoid dehumanizing or add humanizing qualities to them. We live in a time of active proliferation of new and revised management procedures, and designers of information systems are organizational designers as well, who cannot avoid changing organizations (*7, 9*). This is especially true of the proliferation of management information systems, which are more than information systems in the technical sense, as they include all bureaucratic procedures and perhaps all systems components that enter into the production and distribution of goods and services and so dominate the economic, political, and social management of society. Organizational design should be taken on as an explicit activity and management information systems implemented in such a way that they create a more humane setting.

Gouldner (*10*) showed how rules and regulations respond to the self-interest of those who govern and are governed. But to influence the shaping of new bureaucracies and other management systems, it is first necessary to isolate the crucial categories of design features that may make manifest humanizing or dehumanizing qualities of information systems. The analysis present-

The author is director of the Computing Science Program, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6.

ed here is based on the guidelines developed by the Stanley House workshop on humanizing computerized information systems (*11, 12*) in a serious attempt to isolate such design features. The guidelines are grouped into five broad categories, as shown in Table 1.

Many of the Stanley House criteria make sense as procedures for softening a bureaucracy as well as making an information system less rigid. There is no real distinction between manual and automated systems, and guidelines apply whether or not computers are used.

Discussion of Guidelines

By and large, the Stanley House guidelines are self-descriptive. This discussion is designed to illuminate their less obvious aspects and point to special problems that arise in connection with their implementation.

Criterion A2 is not a commonly encountered consideration in systems design. And, indeed, courtesy is not a substitute for real rewards, high quality of service, or other qualities. However, it is possible that courtesy is a prerequisite of humane society. In a rehabilitation hospital where courteous communications were part of a specially designed hospital information system, employees were pleased with that feature and regarded it highly (*8*). It is difficult to evaluate the importance of this courtesy criterion precisely because experience with courtesy in automated systems has been so rare.

Criterion A5 has far-reaching implications for a system's cost and efficiency. One of the favorite methods for optimizing the efficiency and minimizing the cost of a bureaucratic system is to require the individuals being served to supply the necessary information at each procedural component with which they are involved. Further, in order to ensure an uninterrupted flow of work, recipients of service are required to stand in queues at each point. Yet very often the required information can be made available to each procedural component at relatively small cost. It may be particularly important to do this at times when participating individuals are under additional pressures. One pernicious example is the queuing of hospital patients before special treatment or diagnostic centers (such as physical therapy or radiology). Appointments for individual patients made through the hospital information system could eliminate the queues of sick people in drafty corridors so typical of hospital operations. Similarly, a good system could eliminate unnecessary queues and travel by job seekers. On the other side of the coin, we find that the repetitive and unrelieved need to supply a service to queues of recipients is often dehumanizing to service personnel, and the constant demands of the queue prevent trained personnel from applying their skills in a selective manner (*8*).

Criteria A6 and A7 may be related. Large-scale systems tend to be converted onto computers as cheaply as possible. In order to do this a global method of design is often used in which all subprocedures are rigidly defined into a single large structure. The more flexible, albeit much more expensive, way is to build a basic system of linkages to which different procedural modules can be attached. Whenever modifications are required it is then only necessary to reprogram the one affected module. One of the side effects of the global method of design is that it is difficult to modify the system to deal with errors that had not been anticipated. Yet errors of every sort, especially those related to information input, are almost unavoidable in a system that handles a large number of transactions. There is a suspicion in the concerned data processing communities that many corporations leave some errors uncorrected because it is cheaper to lose an occasional customer than to correct for each mistake. The human interface would be a desirable component of a system, even when correction of error may not be the major need. Human contact may be needed for individuals in vulnerable positions, such as the unemployed or the sick, to answer questions about unavoidable delays in providing a service or replying to an application; or just to soften the impact of an impersonal bureaucracy.

The human interface is lacking in most systems we have examined so far, and it may well be that the interface will have to be provided from the outside. One extra-organizational scheme is to have a computer ombudsman serving a large community. Such an ombudsman service could be provided by a professional, consumer, or governmental body, or by a combination of organizations, and would be the mediating link between the perplexed citizen and the perplexing system (*13*).

Related to A6 and A7 is A8, the criterion that management ought to be held responsible for the situation where faulty design causes discomfort and frustration to individuals unable to get relief or attention from a system. Poorly designed systems

Table 1. Stanley House criteria for humanizing information systems.

A. Procedures for dealing with users

1. The language of a system should be easy to understand.
2. Transactions with a system should be courteous.
3. A system should be quick to react.
4. A system should respond quickly to users (if it is unable to resolve its intended procedure).
5. A system should relieve the users of unnecessary chores.
6. A system should provide for human information interface.
7. A system should include provisions for corrections.
8. Management should be held responsible for mismanagement.

B. Procedures for dealing with exceptions

1. A system should recognize as much as possible that it deals with different classes of individuals.
2. A system should recognize that special conditions might occur that could require special actions by it.
3. A system must allow for alternatives in input and processing.
4. A system should give individuals choices on how to deal with it.
5. A procedure must exist to override the system.

C. Action of the system with respect to information

1. There should be provisions to permit individuals to inspect information about themselves.
2. There should be provisions to correct errors.
3. There should be provisions for evaluating information stored in the system.
4. There should be provisions for individuals to add information that they consider important.
5. It should be made known in general what information is stored in systems and what use will be made of that information.

D. The problem of privacy

1. In the design of a system all procedures should be evaluated with respect to both privacy and humanization requirements.
2. The decision to merge information from different files and systems should never occur automatically. Whenever information from one file is made available to another file, it should be examined first for its implications for privacy and humanization.

E. Guidelines for system design having a bearing on ethics

1. A system should not trick or deceive.
2. A system should assist participants and users and not manipulate them.
3. A system should not eliminate opportunities for employment without a careful examination of consequences to other available jobs.
4. System designers should not participate in the creation or maintenance of secret data banks.
5. A system should treat with consideration all individuals who come in contact with it.

are often not corrected because no one is really responsible for their actions. As a consequence, Kafkaesque nightmares may be created for users and participants.

In many ways, procedures for dealing with exceptions may be the most necessary components of a humanized system. The human condition is never so homogeneous that a set of rules can be devised to cover all exigencies. Once bureaucratic procedures are structured, they tend to become rigid even though they may contain provisions to deal with human needs. Exceptions are always difficult to manage. To provide for such flexibility, it is absolutely necessary to provide access to focal points of information or control in order to accommodate a departure from the "norm" where the users' needs require it.

I do not believe that there are technical obstacles to incorporating in working systems the kind of criteria that would permit the consideration of exceptions. My main concern is that obstacles will be generated by unavoidable conflict among humanizing criteria and between such criteria and the use of the system. Consider criterion B1, for example. Some employers of manual job bank programs rely on the face-to-face system to weed out those whom they regard as undesirable applicants. Here is an unstated trade-off between flexibility and equity. Also unstated may be the need to specify whom the system serves. What defines a class of individuals depends, in each case, on the kind of services the system provides or the demands it makes on participants. It is easy to say that a system should at least be aware that affected individuals differ in many personal characteristics and needs and should be accorded correspondingly different types of treatment. However, to achieve that may require an explicit definition of the purposes of a system. For instance, does a job bank serve the job seeker or the employer? It obviously serves the needs of both, and when a conflict exists between these needs it may not be feasible to make that conflict explicit.

In a similar sense conflicts may be created by criterion B4. There is a large variety of situations in which individuals may not wish to avail themselves of services or to provide a system with information touching on their private lives. The whole idea of "choice" is foreign to most large-scale systems, whether automated or manual. The provision of choices may very well mark the border between the dehumanizing and the humanizing system. However, it will add greatly to the complexity of systems, because permitting individual choices may set up conflicts with other criteria or services, including some through which the system seeks to become less dehumanizing. For example, in Canada, Provincial Health Services send an account of services rendered to the head of household. This would seem to fulfill the requirement of keeping the user or recipient of a service informed. Other members of the family, however, might object to finding their health needs reported to the head of household (without necessarily detracting from the affection they might feel for their spouse, parent, or provider). While this problem could be alleviated by addressing the report to the concerned individual, other situations may arise that cannot be easily resolved without providing a wide variety of choices. The spouse of the head of household or the adult children may not wish to inform the head that they have sought medical services. In fact, reporting such information may be harmful to a course of therapy or may needlessly disrupt family life, as when members of a family are seeking treatment for venereal disease or drug addiction, for example.

Opinions are divided about the extent to which information about individuals ought to be withheld from them and from others. Yet there is general agreement that provisions are needed for making access to and evaluation and correction of that information possible.

Criteria concerning actions of the system with respect to information have been widely discussed, so no additional comments may be necessary except in one case—criterion C4. This would make it possible for individuals to add to the system information which they think bears importantly on their background or needs, even if the information is not important for processing their files. This might not add anything to the efficiency of a system, but would add a great deal to the psychological comfort of affected individuals.

Requirements for safeguarding the privacy of individual records may seriously conflict with requirements for humanizing an automated system. In general, the more information a system has about individuals who are affected by it, the more likely it is that it can be humanized, but also the easier it becomes to misuse that information and to violate individual needs or desires for privacy and confidentiality. The extent to which individuals are entitled to privacy or even wish privacy is a matter of political or social decision, as is the extent to which individuals ought not to be dehumanized by a system. Privacy versus humanization is an issue that has not received sufficent attention, and our experience with these concepts is too limited for it to be possible to compare requirements for privacy with those for humanization or make judgments on which is more important. However, it is clear that a very private system with no humane provisions may be just as undesirable as a very humane system with no safeguards to protect the privacy of its participants.

What makes procedural features desirable or undesirable with respect to privacy or humanization can be determined only in the context of the purpose of the system and the safeguards possible. Some systems that list individuals and information about them are desirable and others are not. They may also be desirable and undesirable to different people. For instance, a detailed file on handicapped children in the community would be useful for providing individual services, allocating community resources, and directing planning for schools and recreational facilities. On the other hand, attempts have been made to keep on file the names and records of minors who have been convicted of criminal offenses and to merge such files with other record systems. This has met with opposition from thoughtful members of the community, including members of the police department, and would be very objectionable, at least until adequate safeguards against abuse of such systems have been firmly defined and can be implemented. In the final analysis, it is not only a file's existence but its use which determines its ethical value. Nevertheless, the social and political considerations underlying criterion D1 can be resolved within the context of a particular system. What we are saying is that society can decide whether and how a file of handicapped children or of juvenile offenders should be assembled, maintained, and used.

It may be much more difficult to deal with criterion D2. Central to the problem of privacy is the very much enlarged information base available to government agencies when it becomes possible to merge information from different files. Merging of information may also make many systems more efficient and might make their action more equitable or even more humane. But it may be more to the point that under the guise of humanizing systems or making them more equitable (not necessarily the same thing), the rights of individuals for privacy and freedom from government surveillance in a democratic society may be seriously compromised. For example, the new Insurance Corporation of British Columbia, which is regulated and run by the provincial government, provides compulsory insurance under the name Autoplan for all drivers and car owners and bases its rate structure on records of driver violations. It is disquieting to note the ease with which Auto-

plan has been able to merge court and police files with records of largely business activities of Canadians in British Columbia without a public examination of this important step. Nor has there been public opposition to the extension of Autoplan to other insurance areas. In a similar vein, Lauden (*1*) has shown for four U.S. police and welfare systems how easily information from many sources may be merged. These are perfect examples of the type of activities warned against by Wiener (*5*), who predicted that the needs of large-scale government systems would generate practices which would be discovered only after they were well established.

It is thus clear that the extent to which a system can or will incorporate humanizing or dehumanizing features depends on economic, social, and political decisions. There are limits to the power of managers, engineers, systems designers, and scientists to provide for the inclusion of many desirable features in systems. So we suggest a set of ethical principles—criteria E1 to E5—which, if followed, will ensure that within any set of constraints a system will tend to be humane rather than dehumanizing.

Largely because many transactions of an automated system are difficult to inspect and by their very nature are less open to view than their manual predecessors, the requirement that systems should not deceive or trick, criterion E1, becomes of paramount importance. But even when a system is restrained from deception by law, it may still try to violate the spirit if not the letter of the law. (Common examples are billing practices whereby attempts are made to hide the amount of interest that is being collected from customers or that would be collected if the customer pays only part of what he owes.)

Computerized transactions make it possible for systems to assist participants without needlessly exploiting their labor (criterion E2). The idea that users must provide supportive services in order that a system may function is deeply ingrained not only in the designers of systems but also in the individuals they serve. Members of society are conditioned from birth to stand in line and fill out forms in order to register, to pay, or to receive. They have been habituated to supply information and contribute by their labor wherever they sought to receive a service, were ill, or provided a service for the government (such as paying taxes). It is grotesque but true that when the Nazis led millions of people into concentration camps and eventually into gas chambers, the victims had to stand in lines and deliver their possessions, provide information, and perform all the necessary services required to part them from their goods, their loved ones, and finally their lives. Manual systems burden recipients of a service with a great deal of effort to make the systems function smoothly. Computerized systems do not need to do so, or not really to the same extent. However, the temptation is always there to exploit the willing and conditioned cooperation of members of society. A contrary attitude, that the system should be burdened rather than the human component, needs to be fostered.

Similarly, an attitude should be cultivated by systems designers that all individuals, including employees, who come in contact with a system should be treated with the same consideration (criterion E5). It has been established that organizational structure produces characteristic patterns of alienation. For instance, Blauner (*14*) has shown that workers may develop perceptions of "meaninglessness," "powerlessness," and "work estrangement," depending on how they are fitted into an industry's technology (*15*).

We have chosen to group criterion E3 with ethical rather than economic and social or political considerations. Within the area of information systems and systems control through computers, there are many types of employment that are relatively pleasant and interesting and offer opportunities to large numbers of individuals which are difficult to find elsewhere. The overall cost of eliminating such jobs may be high. This is true when computerization of technology affects jobs that rely heavily on human skills and qualities of perception, attention, and intelligence. There are severe costs when sources of employment that provide interesting, challenging, and above all human types of employment are eliminated. One example of an endangered group, victims of the computerization of communication networks, is telephone operators. It is questionable that replacement jobs for this large number of eliminated positions which offer equally acceptable work for humans are available. The cost of finding employment for the communication workers who ordinarily would have worked for the telephone system has to be borne by society and not by the telephone company, and there is no way to assess or repay the costs to individuals who are forced into less satisfactory employment because opportunities for interesting and humane jobs are eliminated. From an economic point of view, this example shows that a cost-benefit analysis of job elimination through automation should not be based on the effects on a particular industry alone, but should include society as a whole. While it is recognized that it may be difficult for the systems designer to resist the temptation to eliminate such desirable jobs, he should be the first to recognize when they are in danger of being eliminated, and it behooves him as a human being to sound the alarm.

A Final Word About Economics

Perhaps the most serious obstacle to the inclusion of humanizing modules is that they reduce the efficiency of most information systems. Their inclusion will increase overhead in terms of design effort, complexity of procedures, and execution time. It may even be necessary to add to the physical resources of central computers (to provide a larger memory, a greater ratio of input to output, and so on). Consequently, appreciable research along these lines is not expected to be initiated by systems designers and managers, whose primary commitment is to efficiency. While our discussion is not designed to come to grips with the concern of those who are highly cost-conscious, we are nevertheless suspicious of those who refer to humanistic features as negative externalities and who hope that some market mechanism will handle their underlying problem. There is also a "humanistic" side to the debate (*12, 16*).

Lauden (*1*) makes a convincing case that the arrival of the third-generation computer offered new hope for administrative reformers, and indeed many administrative reformers attempted to fulfill this hope almost immediately. The new computer technology promised more closely integrated (which meant centralized) elements of federal, state, and local bureaucracies. It promised better decision-making, better government, better production, better distribution, and better allocation technology. Another important factor, Lauden stressed, is that the value to society of changes in (computerized) information systems does not have to be tested through the electoral process. Similarly, technological changes in industry rarely depend on decisions by stockholders. There are thus factors that shape computerized information systems and restructure means of producing and allocating goods and services or collecting information that are determined solely by political or industrial management and are neither controlled by nor responsive to social pressures. In the case of information systems, political ends are often achieved by management under the guise of instituting cost-saving efficiencies.

The utility of humanizing procedures will not be revealed in ordinary cost-benefit calculations but in the quality of life. Should we burden ourselves and future generations with dehumanizing practices

designed and implemented today? Must not the wish to keep systems humane and dignified take its place with the desire to keep the air and the water palatable as a necessary countermotive to the drive of government and industry to be as efficient and cost-conscious as possible?

Summary

Computerized management information systems increasingly determine all bureaucratic and management procedures that control the production and distribution of goods and services and the collection of information. Thus, they begin to dominate the economic, political, and social management of society. With this domination come procedural features that may dehumanize participants or users affected by the working of most public and private organizations. Yet, despite the overriding importance of a person's dignity and humanity, little is known in terms of scientific specifics about the operational meaning of these concepts or of the antecedent conditions that enhance or diminish them. It will be too late if we wait for knowledge to accumulate about procedures to be incorporated in information systems or information parts of systems to avoid dehumanizing or to add humanizing qualities to them.

A set of guidelines has been developed in a series of workshops sponsored by the Canadian Information Processing Society, Canada Council, and Simon Fraser University. These guidelines may apply where organizational design needs may be met and management information systems implemented in such a way that they create a more humane setting.

References and Notes

1. The most recent analysis for computerized agencies is in K. Lauden, *Computers and Bureaucratic Reform* (Wiley, New York, 1974).
2. For an example of the impact of technology (but not computerization) on an organization see P. Blau, *The Dynamics of Bureaucracy* (Univ. of Chicago Press, Chicago, 1963).
3. J. Weizenbaum, *Science* 176, 609 (1972).
4. T. Sterling, *Humanist in Canada* 25, 2 (1973).
5. N. Wiener, *The Human Use of Human Beings* (Doubleday-Anchor, Garden City, N.Y., 1954); *Science* 131, 1355 (1960).
6. C. Argyris, *Integrating the Individual in the Organization* (Irwin-Dorsey, Georgetown, Ontario, 1965); *Public Admin. Rev.* 33, 253 (May-June 1973); R. Kling, in *Proceedings of the Association for Computing Machinery Computers in the Service of Man* (Association for Computing Machinery, New York, 1973), pp. 387–391.
7. R. Bougeslaw, *The New Utopians* (Prentice-Hall, Englewood Cliffs, N.J., 1965).
8. T. Sterling, S. Pollack, W. Spencer, *Int. J. Biomed. Comput.* 15, 51 (1974).
9. J. Galbraith, *Organizational Design* (Addison-Wesley, Reading, Mass., 1973); T. Whisler, *The Impact of Computer Organizations* (Pergamon, New York, 1970).
10. A. Gouldner, *Patterns of Industrial Bureaucracy* (Free Press, New York, 1954).
11. The guidelines were generated during a number of workshops held at Stanley House, a small estate in the Gaspé at which Canada Council schedules intensive seminars. Canada Council and the Canadian Information Processing Society sponsored one workshop each in 1973. Participating in various of these workshops and otherwise contributing to the formation of these guidelines were R. Ashenhurst, computer scientist (University of Chicago); M. Bockelman, police department (Kansas City); L. Brereton, editor, *Humanist in Canada*; C. Capstick, computer scientist (Guelph University); A. Close, barrister (Law Reform Commission); G. Cunningham, assistant commissioner (Royal Canadian Mounted Police); V. Douglas, psychologist (McGill University); C. Gotlieb, computer scientist (University of Toronto); H. Kalman, historian (University of British Columbia); R. Kling, computer scientist (University of California, Los Angeles); J. Koch, philosopher (Department of Health, Education, and Welfare); P. Lykos, computer scientist (National Science Foundation); S. Pollack, computer scientist (Washington University); H. Schlagintweit, manager (British Columbia Telephone Co.); W. Rogers, provincial auditor (Alberta); D. Seely, computer scientist (Simon Fraser University); M. Shepherd, programmer (Toronto); T. Sterling, computer scientist (Simon Fraser University); and J. Weizenbaum, computer scientist (Massachusetts Institute of Technology). For a detailed description of guidelines see (*12*).
12. T. Sterling, *Commun. ACM* 17 (No. 11), 609 (1974).
13. The Canadian computer ombudsman scheme is developed around a joint effort of the Canadian Information Processing Society and the Consumer Association of Canada (see T. Sterling, *J. CIPS*, in press). The U.S. effort, spearheaded by the Association for Computing Machinery, has as its main concern eliminating an incorrect image of computerized systems and is thus different from the Canadian model.
14. R. Blauner, *Alienation and Freedom* (Univ. of Chicago Press, Chicago, 1964).
15. For confirming evidence of Blauner's findings, see F. C. Mann and L. R. Hoffman, *Automation and the Worker* (Holt, Rinehart & Winston, New York, 1960); C. R. Walker, *Toward the Automatic Factory* (Yale Univ. Press, New Haven, Conn., 1957); A. N. Turner and P. R. Lawrence, *Industrial Jobs and the Worker* (Harvard Univ. Press, Cambridge, Mass., 1965).
16. See, for instance, K. W. M. Kapp, *The Social Costs of Private Enterprise* (Schocken, New York, 1950); E. Richardson, *Work in America* (MIT Press, Cambridge, Mass., 1973).

Software Engineering

A mathematical basis is needed for the practical control of computers in complex applications.

Harlan D. Mills

Computer software began as an afterthought to computer hardware, and, as long as the hardware was small and simple, software could be handled informally by scientifically trained people as a by-product of the use intended for the hardware. As hardware grew in size and complexity, richer software possibilities emerged and software specialists (programmers) appeared, to produce assemblers, compilers, operating systems, and data management systems. Although there was an early recognition of mathematical ideas in computing, for example, in mathematical logic, automata theory, and linguistics, the bulk of these software specialists were pragmatic products of computing practice rather than mathematicians. Thus, although it may seem surprising, the rediscovery of software as a form of mathematics in a deep and literal sense is just now beginning to penetrate university research and teaching, as well as industry and government practices. The forcing factor in this rediscovery has been the growth of software complexity and the inability of informal software practices and management to cope with it.

The author is an IBM Fellow at International Business Machines Corporation, Gaithersburg, Md. 20760. He also teaches computer science part time at the University of Maryland, College Park.

Of course, software makes totally new demands in the sheer volume of logical precision required in its application. A single project may occupy hundreds, even thousands, of people over several years, so that there are unique requirements for recording, communication, and management of the work. These unique requirements lead to almost all of the jargon in software, and, in fact, this jargon tends to obscure the mathematical character of software, as people get caught up in implementation and management details. But the failure to understand this mathematical character led to an overly complex, ad hoc view of software based on historical and accidental ideas, which were often reinvented in ignorance and haste.

The work of Dijkstra and Hoare has been a major force in this rediscovery of software as mathematics. In (*1*, p. 4.2), Dijkstra has presented an argument which sums up the case I want to make here:

> As soon as programming emerges as a battle against unmastered complexity, it is quite natural that one turns to that mental discipline whose main purpose has been since

centuries to apply effective structuring to otherwise unmastered complexity. That mental discipline is more or less familiar to all of us, it is called Mathematics. If we take the existence of the impressive body of Mathematics as the experimental evidence for the opinion that for the human mind the mathematical method is indeed the most effective way to come to grips with complexity, we have no choice any longer: we should reshape our field of programming in such a way that the mathematician's methods become equally applicable to our programming problems, for there are no other means.

This rediscovery of a disciplined basis for software development is finding expression in the recognition of software engineering as a topic for international technical conferences and technical journals. Boehm (*2*) has surveyed software engineering topics from the point of view of the life cycle and management of software. Mills (*3*) has discussed software development.

The Breakdown of Simple Practice

Early computers were scarce and remarkable resources. Thus, meeting the computer on its own terms—reading binary notations, debugging, and patching in machine code, working at 2 a.m.—was a natural thing. Programming was a part of problem-solving with a fantastic new assistant. A little later, programming became a way of dealing with computers on behalf of others—scientists, administrators, or systems engineers. More programming was done on behalf of programmers themselves—building compilers, operating systems, and developing data management systems. But somewhere along the way, the magnificent dreams often turned into nightmares. What seemed easy—the development of a command control system, an information management system, a world-champion automatic chess player—turned out to be extremely difficult or impossible. Major undertakings often took two to ten times as long as had been expected or had to be abandoned.

What happened? The computers had the same simple instruction sets as before. The programmers still knew exactly how the machine worked. The designers still knew how to lay out a system and its development in a systematic modular way, which could be constructed in small units and assembled into larger and larger subsystems until the final system was completed.

What happened was that complexity was growing out of control. Program storage and speed, like dollars, can be measured and aggregated, and budgets set up to manage their allocations and use. But there is no effective concept for measuring program complexity today, and complexity budgeting and control in a quantitative way is unknown. Yet the principal limitation to the development, maintenance, and operation of data-processing systems today is sheer complexity. Hardware storage and speed are more than sufficient; in fact, storage and speed are ill used today because of inefficiencies forced by complexity and overdesign resulting from complexity.

Brooks, the manager of one of the most complex software projects yet attempted, the IBM OS/360 project, noted that "conceptual integrity is *the* most important consideration in system design" (*4*, p. 42) and reported on his experience in managing the development of OS/360 (*4*, pp. 47–48):

> It is a very humbling experience to make a multimillion-dollar mistake, but it is also very memorable. I vividly recall the night we decided how to organize the actual writing of external specifications for OS/360. The manager of architecture, the manager of control program implementation, and I were threshing out the plan, schedule, and division of responsibilities. The architecture manager had 10 good men. He asserted that they could write the specifications and do it right. It would take ten months, three more than the schedule allowed.
>
> The control program manager had 150 men. He asserted that they could prepare the specifications, with the architecture team coordinating; it would be well-done and practical, and he could do it on schedule. Furthermore, if the architecture team did it, his 150 men would sit twiddling their thumbs for ten months.
>
> To this the architecture manager responded that if I gave the control program team the responsibility, the result would not in fact be on time, but would also be three months late, and of much lower quality. I did, and it was. He was right on both counts. Moreover, the lack of conceptual integrity made the system far more costly to build and change, and I would estimate that it added a year to debugging time.

Software Redefined

The term "software" was originally a poetic synonym for "program." But, over time, important reasons have emerged to expand the meaning of "software" to something considerably beyond "program." These reasons are related to a change of view of data processing, from a hardware-centered to a systems-centered perspective. Early computers were such scarce and dramatic resources that it was natural to relate all other activities to them. But, as hardware became more available and embedded in even more complex systems, it often seemed more useful to regard hardware as components in such systems instead of their centers, as discussed by Mills (*5*).

Although such a use was not realized in the early days of computer technology, it is clear now that one major use of software consists of prescribing the harmonious cooperation of asynchronously operating hardware. But, a simple extension is possible here, specifically to incorporate people as agents of action into this formal sphere of harmonious cooperation as well. People have the same architectural characteristics as hardware but with radically different performance parameters in terms of storage, processing, and reliability, and they have radically different instruction sets. A person can carry out the logical operations of a computer (millions of times more slowly), but a person can also execute the instruction "use your common sense" (millions of times faster). For example, in an airline reservation system, customer service people serve as critical analog-to-digital components in converting voice to key stroke and digital displays back to voice in dealing with customers.

There is another, older use of software, to directly support a specific piece of hardware, which makes that device easier to use either by people or by programs. For example, a program used to control a tape drive may take in commands to read, write, or rewind the tape, but put out instructions to motors or to read-write heads at a much lower level of detail; when errors are detected, the program may call for automatic retry of physical operations, all unknown to the program user. In short, such software combined with the hardware makes up an "abstract machine," easier to use as a component than the hardware itself. For this reason, I will use the term "machine" to mean any command-driven combination of hardware and software, not just base hardware.

A set of such hardware-software machines can, in turn, be organized, through additional software which provides for their harmonious cooperation, into a larger, more capable, abstract machine, which a user (person or program) can command and receive results from as a single agent. For example, a simple computer, with its card readers, printers, and other components, all operating asynchronously under the control of an operating system, can be viewed as just such an abstract machine.

Finally, people can serve as agents of action in abstract machines, in much the same way as the hardware in the foregoing discussion, if the term "software" is expanded to include the ideas

of user's guides and operating instructions for people in their specific roles in a system.

For these reasons, software is better defined as "logical doctrine for the harmonious cooperation of people and machines." Such doctrine takes the form of programs for hardware, and combinations of user's guides and operating instructions for people. The programs may be microprograms, stored in read-only storage (for example, "firmware"), as well as ordinary programs. In brief, software defines a system of abstract machines, some of which call on other abstract machines, until people and hardware are reached as the ultimate agents of action in the system.

Data Processing

The data-processing explosion. Although nothing of the kind existed 25 years ago, every sizable business, government agency, and institution in the United States now depends in a critical way on large data-processing systems. Such a system often interfaces thousands of people with a complex of machines—computers, terminals, data-entry devices, many of them in real-time communications over considerable distances. These large systems may have large central computing subsystems, or the computing may be distributed over several centers. They are inextricably tied in with the operations of the enterprise, involved in an astonishing variety of science, engineering, and commercial tasks.

With such a short history, it is surprising that these large data-processing systems exist at all. And yet they are no gleam in a crystal ball. They are here now, and working well enough to be indispensable to every sizable undertaking in the country but working poorly enough to frustrate the managements and beneficiaries of these undertakings also.

The engineering of computer hardware developed out of the scientific research that made it possible, and manufacturers of data-processing hardware benefited directly from that science and its engineering. But the problems of putting these products of science into practical use filter back to the research community much more slowly, because the applications are not usually scientific but are instead industrial, financial, or administrative. That situation in data processing is accentuated by its very rate of development. The close alliance of science and engineering in hardware has over the last 25 years accomplished dramatic achievements in reliability, speed, storage, cost, and function. This achievement has few parallels in human history.

But the uses of this remarkable hardware in industry and government have led to large, complex, ponderous, unreliable, incompatible data-processing systems with no significant precedents or guides from science, such as helped to shape the hardware. The precedents for such complex systems are in the day-by-day operations of business and government, as conducted by people, with all the local variations and considerable inconsistencies that only people could tolerate in operational activities. As a result, during the past 25 years there has been a much slower rate of improvement, however it may be measured, in the use of hardware than in the hardware itself.

The growth of data processing in industry and government reflects the increased complexity, interdependency, and competition among enterprises. One example of the interdependency is the reporting by industry of payroll information to the Social Security Administration and the Internal Revenue Service. The information is transferred routinely in computer format (this procedure is based on the presupposition of widespread computer operations). The competition to provide better services to customers through data processing is apparent in hotel and airline reservation systems, insurance and banking, and retailing. The competition that generates these systems also creates a demand for their improvement—to systems that are more capable, more reliable, more economical, more manageable. In short, there is a remarkable hunger across both industry and government for better ideas and practices from science and engineering in data processing.

Why data-processing practices are backward. The data-processing industry is a spectacular product of science and engineering. But its practice leaves much to be desired. It is suffering the growing pains of an infant industry that has had no time to develop, try out, and select sound industrial practices—it is a mixture of great wisdom and foolish folklore. It needs help from science and engineering in the worst way. But the help it needs is not simple to provide.

The bulk of managers and workers in data processing have moved into the field laterally, without university education in computer science or software engineering. The principal sources of information for these people are hardware manufacturers and trade publications. Professional societies and journals play an indirect role in transmitting scientific and engineering ideas to this community. But there is little help to be found in trade publications of the industry, because the problems are of a deep and systematic nature, of unprecedented logical complexity, and so patent remedies are of little value.

In the early days of computing, many very good people from the nation's great universities and laboratories moved into the field and brought with them or developed many sound and enduring ideas. With the growth of data processing, the bulk of people coming into the field later were less highly motivated and educated. As a result, data processing has expanded in quantity but has decreased in average quality, in personnel, over its short history.

This lag in industry practices is of particular concern when one is endeavoring to identify research needs. Research efforts take years of educational preparation and then deep concentration, often over decades. But in data processing, with only a 25-year history, industry needs are just beginning to emerge. What seemed important 10 or 15 years ago may be less important now because of this changing set of needs. Thus it is all the more critical to properly assess future trends in data-processing practices, in order to anticipate research areas of more lasting relevance.

International competition in data processing. Data processing has come to be an important national asset in the management and organization of industrial resources in all well-developed countries. The value of this data processing depends on the quality of software found in industry and government. For example, if poor or unreliable software in a grocery chain contributes to the spoilage of foodstuffs shipped to the wrong place, the energy required to grow the food is wasted. Thus, the cumulative effect of software quality has a significant impact on the relative industrial (and social) health of each country. If this is so in the first 25 years, it will be even more the case over the next 25 years.

Because of this importance of data processing in national affairs, the following incident in space activities teaches us a lesson. In the late 1950's the United States was behind the U.S.S.R. in missile booster technology (because it was ahead in the miniaturization of atomic bombs), and this fact was a cause of considerable consternation in the United States. In retrospect, being behind in

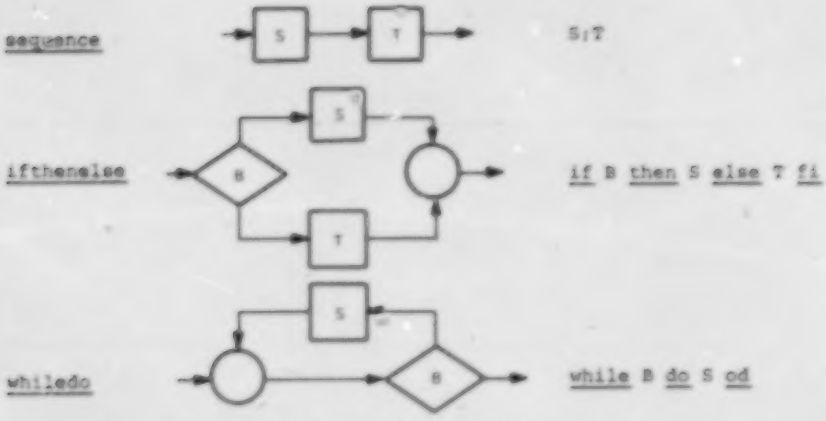


Fig. 1. A set of action primitives.

booster technology did the U.S. electronics industry a real service, because it imposed a discipline for miniaturization that would otherwise have been difficult to enforce. In data processing the situation in the two countries is reversed. For the last 20 years the United States has maintained a commanding lead in computer power, computers being scarce resources in the U.S.S.R. The result is that the U.S.S.R. now has fewer, but more highly disciplined programmers than the United States.

It is already clear that Japan will be an important international force in data processing. In addition to Japan's already proven ability in electronics production, it has two specific advantages for software development in the future (which are also shared by the U.S.S.R.). First, the Japanese government has the means and apparent resolve to organize and coordinate software research and development as an instrument of national policy, as evidenced by its formation of a Joint System Development Corporation (6). Second, Japanese industrial and labor practices will permit a smooth transition into the automation of industrial processes, by comparison with the case in the United States where automation is often regarded as a threat to jobs and labor and subject to political and social delay and disruption.

The Emerging Idea of Software Engineering

Software tools—technical proving grounds. In spite of the complexity barrier, data-processing systems of large size are in widespread operation. In most cases they have evolved over decades or more into their present forms, and they usually represent more perspiration than inspiration in their present forms. The major tools for their development and maintenance are compilers (and assemblers) for programming languages and operating systems (often with extensive data management facilities). The development of these tools has been a major basis for technical development in programming itself. Programmers have a unique opportunity to build their own tools. But, with only a few exceptions, the tools themselves have provided more experience in what to avoid next time than as examples of well-designed and documented software.

The development of compilers presents deep and interesting problems in programming. It involves text and language processing, which has motivated considerable research in formal grammars and their languages, and studies in finite state machines. The compiling process calls for the management of extemporaneous data storage and for the accessing of data in flexible ways. Usually, the problems of programming compilers are more difficult than the programming problems solved in the languages they compile; for example, a programming language with very simple data storage facilities will need a compiler which itself requires much more flexible data storage.

The idea of an operating system has evolved dramatically over the past two decades, both to adapt to new hardware capabilities and to respond to new human demands. Human interfaces have expanded from a single operator entering and reading binary coded numbers at a primitive console to hundreds of users in simultaneous conversation with friendly terminals. Operating systems deal with multiple processors and multiple programs in concurrent execution in these various processors, accepting spontaneous, asynchronous signals and data between themselves and from outside, automatically shuttling programs and data to and from high-speed and low-speed storage devices between bursts of execution. The development of software to supervise this complex asynchronous activity has been a second major challenge in software development and a source of much research in asynchronous program control, data management, and large system decomposition techniques. As in the case of compilers, the applications systems which run in these operating systems are usually of simpler construction, with less demanding logical and performance requirements, than the operating systems.

In short, compilers and operating systems have served as important proving grounds for programming techniques. The ideas emerging from these proving grounds invariably find use in applications systems. For example, the use of formal grammars and languages, inspired by compiler problems, is now widely recognized in the design of applications interfaces with users; techniques of large system decomposition originating in operating system designs are now seen to be useful in applications systems as well.

Three software discoveries. The first computers were single sequential machines which were operated under program control. But, curiously enough, before people learned how to program these simple machines adequately, new hardware was invented which was capable of cooperative asynchronous operation, and so people immediately set out to program the new hardware before they had time to master the old. In retrospect, it became necessary, mathematically, to retrace these steps—to study the programming of sequential processes and then the programming of the harmonious cooperation of independent sequential processes. In the past decade, this study has led to three significant software discoveries: (i) the dependable design of correct sequential processes; (ii) the synchronization of independent sequential processes for harmonious cooperation; and (iii) the organization of software-hardware into systems of abstract machines.

Dijkstra has played a leading role in bringing all three of these discoveries to light. In the first place, he introduced the idea of "structured programming" (7) as a systematic method of design for sequential processes, based on the need for mathematical proofs for their correctness. In order to cope with the scale-up in size, in order to keep proof size proportional to program size, he found it necessary (and, surprisingly, possible) to restrict the freedom of design to a small,

finite set of primitive sequential processes. This restriction was described in a famous letter to the editor entitled "GOTO statements considered harmful" (8) in 1968, which initiated considerable controversy but which time and experience have pretty well settled.

Structured programming. In spite of the controversy over GOTO's, the essence of structured programming is not the absence of GOTO's in programs but the presence of rigor in programming. This rigor is achieved by a so-called stepwise refinement process, which expands an action specification into a primitive sequential process of (smaller) action specifications, carried out repeatedly until actions of the programming language are at hand. Thus, beginning with an initial action specification for a program as a whole, the program is designed, stepwise, into a hierarchy of refinements.

A set of such primitives, given in flow chart and text form, is shown in Fig. 1, where S, T are actions (which may be further expanded in these same primitives) and B is a (Boolean) test. In illustration, after several expansions, a sequential process might be defined by the program, in flowchart and text, shown in Fig. 2, where the text form has been broken into lines and indented to show the structure of the expansions. It should be added that the foregoing gives the form but not the substance of good design. A good design not only conforms but satisfies many additional criteria not easily stated here. But the form does keep the correctness proof size proportional to the program size.

The principal reason for the controversy about structured programming was the question of whether such a small set of simple primitives was adequate to achieve sequential control. At the time, there was little discipline in such control logic, because it was widely supposed that complexity was a necessity of good design. We now know better.

Structured programming has already found important practical applications in software development, as discussed by Baker (9). There is already a widespread acceptance of the idea in industry and government, even though practice often lags behind intentions.

An axiomatic basis for sequential processes. Hoare (10) provided an explicit confirmation of Dijkstra's objectives in structured programming by providing an axiomatic basis for proving the correctness of each of the primitive sequential processes proposed by Dijkstra. Thus, at each stepwise refinement, a standard procedure can be invoked, at any level of formality desired or appropriate. Specifically, Hoare introduced the idea of a "precondition" P, and "postcondition" Q for an action S, written $P\ \{S\}\ Q$, and interpreted "if P holds and sequential action S is carried out (and terminates), then Q will hold." Now, the primitive sequential processes above can be axiomatized as follows:

sequence

$$P\ \{S\}\ Q \wedge Q\ \{T\}\ R \Rightarrow P\ \{S;T\}\ R$$

ifthenelse

$$P \wedge B\ \{S\}\ R \wedge P \wedge \neg B\ \{T\}\ R \Rightarrow P\ \{\textit{if } B \textit{ then } S \textit{ else } T \textit{ fi}\}\ R$$

whiledo

$$P \wedge B\ \{S\}\ P \Rightarrow P\ \{\textit{while } B \textit{ do } S \textit{ od}\}\ P \wedge \neg B$$

The last axiom, for the *whiledo*, is especially illuminating, because it identifies a condition P, called an "invariant" (by reason of the persistence of P in the antecedent), which can be used to prove looping programs correct. It is now easy to see that Dijkstra's claim for the proof sizes of structured programs is correct—the form of the proof is independent of the depth of the hierarchy at each refinement step. The practical possibilities of proof of correctness have led in turn to programming language ideas which describe not only code but proof, as shown by Wegbreit (11).

More recently, Dijkstra (12) has used the foregoing ideas of Hoare to introduce the concept of "predicate transformers" for sequential process design. In brief, for each Q, S defines a "weakest precondition" P, such that $P\ \{S\}\ Q$, so that S can be interpreted as a predicate transformer (which transforms Q into P). Such a weakest precondition always exists, because, regarding P, Q as predicate valued actions, the sequence $S;Q$ is, in fact, the weakest precondition P for which $P\ \{S\}\ Q$. Now, the condition required of a sequential process must be known in order to design it, but this is postcondition Q. So the sequential action S which transforms Q into a suitable predicate P can be determined as a stepwise refinement. Dijkstra showed how each primitive sequential process behaves as a predicate transformer and thus laid the groundwork for a method for deriving programs which are automatically correct (12). For example, denoting the predicate transformer *wp* (weakest precondition),

$$P = wp(S, Q)$$

then the transformer for the sequence $S;T$ is

$$wp(\text{"}S;T\text{"}, Q) = wp(S, wp(T, Q))$$

Gries (13) has given an elementary exposition of the use of predicate transformers.

Structured programs and function algebras. A sequential process S defines a "program function"

$$s = \{(x, y) \mid y = S(x)\}$$

and a Boolean test B defines a predicate function

$$b = \{(x, y) \mid y = B(x)\}$$

The primitive sequential processes axiomatized above can be identified with operations in an algebra of functions, namely,

sequence operation

$$s;t = \{(x, y) \mid \exists z\ (z = s(x) \wedge y = t(z))\}$$

ifthenelse operation

$$\textit{if } b \textit{ then } s \textit{ else } t \textit{ fi} = \{(x, y) \mid (b(x) \wedge y = s(x)) \vee (\neg b(x) \wedge y = t(x))\}$$

whiledo operation

$$\textit{while } b \textit{ do } s \textit{ od} = \{(x, y) \mid \exists k \geq 0\ (\forall j, 0 \leq j < k,\ (b(s^j(x))) \wedge y = s^k(x) \wedge \neg b(y))\}$$

(in the *whiledo*, k is an "iteration count," b must test true up to $k - 1$, y must be the result of k iterations of s, and

```
if B1
  then
    S1;
    while B2
      do
        S2
      od
  else
    if B3
      then
        S3
      else
        S4
    fi;
    S5
fi
```

Fig. 2. A simple structured program.

b must test false to terminate). Here, the sequential control symbols; *if, then, . . . , od* are *infix, prefix, postfix* operators (just like +, −, and other symbols in ordinary algebras), with functions as operands and results. A structured program corresponds to a compound expression in such an algebra of functions, and the various convenient properties of hierarchical structure, indentation capabilities, and subprogram nesting are all seen to be a result of this algebraic structure.

The stepwise refinement process can be formulated as solving function equations, one of

$$r = s;t$$
$$r = \textit{if } b \textit{ then } s \textit{ else } t \textit{ fi}$$
$$r = \textit{while } b \textit{ do } s \textit{ od}$$

in which r is given and b, s, t are to be determined. Closed-form solutions can be given for each of these equations, as shown by Mills (*14*). For example, the solution for the *whiledo* equation has a classical mathematical form—an existence condition (on r), and a single parameter family of solutions (with a function as the parameter). Reynolds and Yeh (*15*) give another view of programs which compute functions by induction.

Process synchronization. Each hardware device in a computing system operates asynchronously with every other unit, so that signaling and waiting are practical necessities in their coordination, that is, their harmonious cooperation as sequential processors. In this way card readers, printers, tape units, and logical processors can all cooperate in a single coherent operation, even though each device operates at quite different and mutually unpredictable rates. For example, a central computer can request action from a card reader, continue some calculations, and be interrupted when the card reader has completed its operation.

Dijkstra showed how the necessity of programming asynchronous processors could be converted into a powerful software design technique, by organizing an entire system as a "society of sequential processes, progressing with undefined speed ratios" (*16*, p. 343). Such sequential processors are not tied to specific hardware devices (although each asynchronously operating device defines one or more sequential processes in a natural way) but are used to organize (divide and conquer) more complex processing activity into a set of simpler ones which are coupled through synchronizing primitives called "semaphores." Semaphores play two distinct roles: (i) the mutual exclusion of sequential processes when any one of them is performing a critical task, for example, updating a commonly used record, and (ii) providing synchronization signals to one another. The use of semaphores permits proof of cooperation among sequential processes, such as guaranteed progress (absence of deadlock) to goals and integrity of data common to the processes, as shown by Dijkstra (*17*), Habermann (*18*), and Hoare (*19*).

Brinch Hansen (*20*) and Hoare (*21*) have further developed the synchronization of sequential processes through program units called "monitors," which led to methods for extending Hoare's axiomatic treatment of sequential processes to axioms for the synchronization and exclusion of cooperating processes. Howard (*22*) and Owicki and Gries (*23*) have expanded on this treatment.

Abstract machines. The organization of a sequential process into a hierarchical structure through stepwise refinement is a powerful abstraction. The additional synchronization of several sequential processes into a single coherent superprocess is even more powerful and brings the ability to organize any system of people and hardware into harmonious cooperation. First, however, let us consider only hardware, in order to develop the idea more simply.

A number of investigators, beginning with Parnas (*24*), Liskov and Zilles (*25*), and Wulf *et al.* (*26*), have observed that the organization of software systems can be improved in clarity and reliability if certain data and programs are organized as "data abstractions," namely, as software-defined finite state machines, which are "black boxes" to their users. For example, such an abstraction could behave as a first in, first out queue through a set of external commands, with the internal organization of the queue hidden from the user. An advantage of this arrangement is that the queuing mechanism can be entirely changed without affecting the rest of the programs that make use of it. Another useful and common abstraction is a card reader, a software-hardware machine which operates asynchronously with a using process.

In fact, I will use the term "abstract machine" to mean any software-hardware-defined finite state machine. The implementation of an abstract machine is done in only one way—by means of a society of cooperating sequential processes, each of which may employ abstract machines in turn. Now it is easy to see that the recursive use of abstract machines allows us to explain and design the largest of software systems. But, first, there is a lesson in hardware machines.

A well-designed and documented hardware machine has the following properties, as noted by Habermann *et al.* (*27*) in general: (i) a full specification of its legal (meaningful) operations, (ii) the detection of illegal requests for operations and a description of the response taken to them, and (iii) the guarantee that illegal requests do not damage the machine, that is, do not prevent the machine from operating correctly [as defined by properties (i) and (ii)] if restarted. Furthermore, even though the timing specifications of operations may be only approximate (as a result, for example, of mechanical variances), at the completion of any logical operation the new state of the machine should be specified precisely. We should expect no less from an abstract machine.

The idea of abstract machines makes possible the development of hierarchies and levels of abstractions in software design. A software system can be designed as a set of harmoniously cooperating sequential processes, which have powerful abstract machines at their disposal. Each of these abstract machines can in turn be designed as a new set of harmoniously cooperating sequential processes, with simpler abstract machines being used until hardware machines are reached. For example, an airlines reservation software-hardware system can be designed as a cooperating set of abstract terminals, an abstract control program, and an abstract data management system. An abstract terminal, in turn, can be designed as a combination of a cooperating abstract keyboard, an abstract display, and abstract local data. This whole software-hardware system then becomes itself an abstract machine for a system of people and machines that can deal with the public.

References

1. E. W. Dijkstra, *On a Methodology of Design* (MC-25 Informatica Symposium), Mathematical Centre Tracts, Amsterdam, 1971).
2. B. W. Boehm, *IEEE Trans. Comput.* **C-25** (No. 12), 1226 (1976).
3. H. D. Mills, *IEEE Trans. Software Eng.* **SE-2** (No. 4), 265 (1976).
4. F. P. Brooks, *The Mythical Man-Month: Essays on Software Engineering* (Addison-Wesley, Reading, Mass., 1975).
5. H. D. Mills, *On the Development of Systems of People and Machines* (Lecture Notes in Computer Science: 23, Programming Methodology, Springer-Verlag, Berlin, 1975).
6. E. K. Yasaki and A. Pantages, *Datamation* **22** (No. 9), 91 (1976).
7. O. J. Dahl *et al.*, *Structured Programming* (Academic Press, New York, 1972).
8. E. W. Dijkstra, *Commun. Assoc. Comput. Mach.* **11** (No. 3), 147 (1968).
9. F. T. Baker, *Assoc. Comput. Mach. SIGPLAN Notic.* **10**, 172 (June 1975).

10. C. A. R. Hoare, *Commun. Assoc. Comput. Mach.* **12** (No. 10), 576 (1970).
11. B. Wegbreit, *IEEE Trans. Software Eng.*, in press.
12. E. W. Dijkstra, *A Discipline of Programming* (Prentice-Hall, Englewood Cliffs, N.J., 1976).
13. D. Gries, *IEEE Trans. Software Eng.* **SE-2** (No. 4), 238 (1976).
14. H. D. Mills, *Commun. Assoc. Comput. Mach.* **18** (No. 1), 43 (1975).
15. C. Reynolds and R. T. Yeh, *IEEE Trans. Software Eng.* **SE-2** (No. 4), 244 (1976).
16. E. W. Dijkstra, *Commun. Assoc. Comput. Mach.* **11** (No. 5), 341 (1968).
17. ———, in *Programming Languages*, F. Genuys, Ed. (Academic Press, New York, 1968).
18. A. N. Habermann, *Commun. Assoc. Comput. Mach.* **15** (No. 3), 171 (1972).
19. C. A. R. Hoare, in *Operating Systems Techniques*, C. A. R. Hoare and R. H. Perott, Eds. (Academic Press, New York, 1973).
20. P. Brinch Hansen, *Assoc. Comput. Mach. Comput. Surv.* **4** (No. 4), 223 (1973).
21. C. A. R. Hoare, *Commun. Assoc. Comput. Mach.* **17** (No. 10), 548 (1974); *ibid.* **18** (No. 2), 95 (1975).
22. J. H. Howard, *ibid.* **19** (No. 5), 273 (1976).
23. S. Owicki and D. Gries, *ibid.*, p. 279.
24. D. L. Parnas, *ibid.* **15** (No. 5), 330 (1972).
25. B. H. Liskov and S. Zilles, *IEEE Trans. Software Eng.* **SE-1** (No. 1), 7 (1975).
26. W. A. Wulf, R. L. London, M. Shaw, *ibid.* **SE-2** (No. 4), 253 (1976).
27. A. N. Habermann, L. Flon, L. Cooprider, *Commun. Assoc. Comput. Mach.* **19** (No. 5), 266 (1976).

Human Performance Considerations in Complex Systems

H. O. Holt and F. L. Stevenson

Acceptable human performance in complex systems depends upon precise human-machine interaction. Such interaction is the focus of attention of design engineers and computer programmers, for example, on the one hand, and of human performance psychologists on the other. The meaning and extent of that interaction has evolved and expanded over the years. We now commonly find computers of various sizes on the machine side of the human-machine interface, and their presence has changed human performance considerations markedly. On the human side we have seen an accelerating emphasis upon people as information processors, thus adding many considerations to the older—but persistent—anthropomorphic concerns.

In this article we review the "human factors engineering" field briefly, and then discuss in some detail the requirements that computers have put on people and human performance technology in computer-based systems. For examples, and in the citation of solutions, we draw heavily upon our own experience in the Bell System.

There have been human-machine interaction concerns of a sort ever since primitive man first extended his own abilities with simple weapons and tools. In more recent history the industrial revolution accelerated greatly the transfer of work functions from people to machines and complicated the human-machine interface problems considerably.

Dr. Holt is director of the Human Performance and Support Center, and Mr. Stevenson is head of the Systems Training Department at Bell Laboratories, Piscataway, New Jersey 08854.

Human Factors Engineering

It is generally agreed that World War II marked the beginning of a professional approach to what came to be called human factors engineering, that is, a systematic approach to studying problems of human-machine interaction and to arriving at practical solutions on a scientific basis. Before the war, going back into the late 19th century, systematic work had been done by psychologists, but it tended to focus upon selecting or training people to interact with machines. But the tremendous industrial and military expansion brought on by World War II, and the greatly increased complexity of the weapons systems being produced, complicated human-machine interaction considerably, so that the selection and training approach no longer was sufficient. For example, it was a simple matter to get a relatively small number of men to fly the slow, uncomplicated fighter aircraft of World War I as compared to getting thousands of men to perform satisfactorily in the high performance P-38s and P-51s of World War II. Thus, a recognized professional specialty, usually known as human factors psychology or human engineering, was spawned.

Since human factors psychology came into being during World War II, it is to be expected that it would continue to thrive in a military environment after the end of that war. The Army, Navy, and Air Force all established substantial centers for the study and application of this discipline, and many of them still exist today. Human factors practitioners spread into the industries that supplied military equipment and systems. There was a particularly large concentration in the aerospace industry. The movement also took hold in many nonmilitary fields such as transportation, telephony, and occupational safety.

In 1953, Paul Fitts and others typified the work being done in the 1940's and 1950's as follows: "In the design of equipment, human engineering places major emphasis upon efficiency as measured by speed and accuracy of human performance in the use of the equipment. Allied with efficiency are the safety and comfort of the operator. The successful design of equipment for human use requires consideration of the man's basic characteristics, among them his sensory capacities, his muscular strength and coordination, his body dimensions, his perception and judgment, his native skills, his capacity for learning new skills, his optimum work load, and his basic requirements for comfort, safety, and freedom from environment stress" (*1*).

Thus traditional human factors engineering concerns itself with data gathering and experimentation meant to yield precise information about human capabilities. With such information, machines can be built to fit humans. For example, studies revealed design requirements for visual displays of understandable information so that correct decisions or control actions can be made without delay. Comfortable physical fit between people and machines can be established with the use of such information as the average human's physical size, strength, and reach. This information has been stored in handbooks for the use of equipment designers and for human factors personnel who work with equipment designers.

Personnel Subsystem Development

By 1960 there was a growing recognition that profound changes needed to be made in the organization and execution of complex system designs. In brief, an opinion was growing that there should be a simultaneous, integrated design and development of the hardware and software, and of the human elements of the system. In 1960 Van Cott wrote the following: "Traditionally, the goal of human engineering has been to engineer machines for human use—to make equipment compatible with man's characteristics and limitations. In this book, however, it is assumed that the primary goal of human factors engineering is to help design an optimal system. The human factors program helps design not only the equipment, but also the jobs, and to some extent the characteristics of the personnel in the system through appropriate selection and training programs. This approach changes the typical role of the human engineer from that of a design critic to that of a component specialist and a member of the system design team" (*2*).

By 1960, too, not only was there a healthy appreciation of complex weapons systems and the important role of human factors engineering in their designs, but the importance of the computer was being realized and there was a rising concern about the complexity of the interaction between humans and computers. In that year, an article by Licklider appeared in the first issue of the *IRE Transactions* on "Human Factors in Electronics," in which he raised three issues of "man-computer symbiosis": the language mismatch between computers and people, the physical interface (computer console and terminals), and the speed-cost mismatch between human and computer (*3*).

In 1977 we no longer hear much said about Licklider's three problem areas, at least not expressed in the same terms. We have been able to compensate for the speed-cost mismatch between man and computer by such approaches as time-sharing. In information systems, our concern often is the opposite: the response time of a computer to an operator inquiry sometimes is not fast enough for optimum system performance.

With regard to the physical interface, the variety of available well-designed terminals is much better than it was a few years ago. This improvement may be due at least partly to the negative reactions to many of the early terminals which were so poorly designed.

The language mismatch between people and computers—at least one aspect of it—still is a major problem. The matter of preparing and debugging programs, especially for large software projects, still is, as one authority puts it, "in the category of a cottage craft" (*4*).

But today there is another major problem of "human-computer symbiosis." It concerns the fit between the computer and its software (computer subsystem) and the people (personnel subsystem) in the same system. The term "personnel subsystem" refers to the manual procedures, the human-computer interface, training, documentation, and performance aids.

Today's problem and its solution are familiar to every human factors psychologist who has worked in a large human-computer development. The software is designed, written, and debugged with meticulous care, while the personnel subsystem is left to the last minute and is given little or no consideration. In the worst cases the software simply is turned over to the clients who attempt to fit in the people functions some way—often with disastrous results. (Have you ever attempted to get an error corrected in a computer-based charge account?) In less traumatic cases, the software designers, with or without the help of human factors people, attempt to adapt the human functions to the computer functions, which may allow the system to run, but often in a suboptimum manner. The correct way to proceed, according to Van Cott and others, is with a simultaneous, integrated design of both subsystems.

We will now consider the issues raised by Van Cott and Licklider, and give illustrations from our own experiences of how these issues have been addressed in the development of computer-based systems, particularly in the past 10 years. To repeat the issues: Van Cott said ". . . it is assumed that the primary goal of human factors engineering is to help design an optimal system. . . . This approach changes the . . . role of the human engineer from that of a design critic to that of a component specialist and a member of the system design team" (*2*). Licklider raised the issue of human-computer symbiosis, which we interpret to include the difficult task of coordinating human information processing and computer data processing into an integrated system.

System developers have addressed the issue of human-computer symbiosis in varying ways. One approach has been to pay only casual attention to human factors and personnel subsystem considerations. A second approach has been to devote extensive resources to the development of training courses for the personnel. A third approach has been to concentrate on the human engineering of the human-computer interface. The fourth approach has been to design the complete personnel subsystem concurrently with the computer subsystem. We will discuss each of these approaches separately.

Develop Hardware and Software, but Very Little Personnel Subsystem

Many new systems involve the application of a new technology to existing business or scientific functions. Making full use of a new technology—for example, computer technology—is a challenging undertaking. Development organizations sometimes become preoccupied with applying the new technology; all their attention and resources are devoted to it, and little or no attention is given to the personnel subsystem.

One consequence of such an approach is that human errors often become intolerable. This can be illustrated by an early instance of computerization. The plan was to mechanize part of a customer records and billing process. Hardware was selected and software developed. Soon after installation, the system became swamped with errors. Eventually, and with great difficulty, improvisations were made in the personnel subsystem which brought the error rate within tolerable limits.

The worst problems are traditionally with input errors. Computer edits identify errors and reject the error-containing transactions. Error messages are then fed back to personnel in the error-correction feedback loop. There is a snowballing effect: "temporary" personnel are added to correct errors; system performance keeps dropping; the rate of system throughput drops because most of the data processing effort is expended in the error-correction process.

Abnormally high error rates have three consequences. First, if outputs are delayed, customer schedules (and customers) are upset. Second, the cost of extra processing required to correct errors can be high; for example, a study of errors on customer service orders showed the total cost to correct each error to be $7.25. Such costs can wipe out the savings which the new system was to effect. Third, high error rates demoralize system users. We once asked a clerk whose sole job all day long was to prepare a complex input form, how many forms were returned for correction of er-

rors. The answer: "All of them, and some more than once."

Another consequence of the single-technology approach is that system users often are unable or unwilling to use system features which have been devised at great expense. Licklider, describing the SAGE (semi-automatic ground environment) system, reported: "According to credible reports, there were makeshift plastic overlays on the cathode-ray displays, and the scope watchers were bypassing the elaborate electronics—operating more or less in the same 'manual mode' used in World War II" (5).

Fig. 1. Directory assistance operator working at specially designed terminal of experimental computer-based system.

Produce Extensive Training Course Materials

System designers, when confronted with widespread human performance problems in their systems, search for a broader design approach. They sometimes believe that training equates to satisfactory human performance. There are at least two flaws in such an assumption. First, human performance in systems is the complex interaction of many variables, only one of which is training. Fox, citing experience with developmental testing, described eight major variables affecting human performance in systems; only one was training (6). (The other variables are: design, human-machine interface, information transfer, environment, personal factors, supervision, and documentation.)

Second, simplistic as it may sound, system personnel should be trained to do the manual procedures of the system. However, the most carefully developed training courses cannot compensate for a poorly designed or suboptimized personnel subsystem. The procedures themselves may be incorrect; for example, some information should be stored in performance aids instead of memorized. No matter how much training they receive, system personnel probably will be unable to overcome many of the error-producing procedural and interface problems inadvertently incorporated in the personnel subsystem design.

During the 1970's, instructional technology has become widely used. This method of training course development features thorough analysis of tasks to be taught. We have found that a task analysis, done for training purposes, often reveals personnel subsystem design deficiencies. It also often reveals the need for new performance aids and procedural documents. The net effect usually is reduced training time.

For example, we were asked to help with training course development for a new system which provides standardized circuit designs. A task analysis revealed the need to add performance aids and restructure user documentation. The result was a 30 percent reduction in training costs, improved human performance, and a reduction in the volume of system documentation that was required.

Design the Human-Machine Interface

Systems designers are often familiar with some of the terminology and concepts of human factors engineering; therefore, the "human factoring" of interfaces has traditionally received support in both large and smaller systems.

One such project was MECHSIM (mechanical simulation of a computer-based directory assistance system). This project had as its ultimate goal the reduction of the amount of time an operator spends responding to a request for directory assistance. The MECHSIM study followed some 10 years of low budget research and experimentation directed to reducing operator work time.

Control or reduction of operator work time is important to telephone companies. Directory assistance is expensive to provide and, until recently has been provided universally at no charge. In 1970, for instance, the Bell companies had more than 44,000 people working in directory assistance offices at an annual cost of $300 million. The average duration of a directory assistance call is about 33 seconds; therefore, the annual value, so to speak, of each second is in excess of $9 million. So, if Bell laboratories could show the telephone companies a way to cut a few seconds off average call duration, the savings would be considerable.

Operators on the job quickly get quite proficient at using the traditional "paper data base," that is, telephone directories, thus early attempts to improve the process were unsuccessful. For example, in 1956, a trial was run in which directory pages were put in microfilm "sticks," as they were called, and projected for the operator on a ground glass screen. This method did not save look-up time, largely because the state of the arts of optics and microfilm were not sufficiently advanced.

In the early 1960's, analytical studies were undertaken with the view to utilizing the computer and its related cathode-ray tube (CRT) terminal. A major question was, how would the operator gain access to the file if it were in the computer instead of on paper? When operators used paper telephone directories, access was by full name of the customer; however, the studies showed that if the operator were to key in trigrams—that is, the first three letters of last names instead of the entire name, input time and spelling and key stroke errors would be reduced.

In 1965, a human factors trial was instituted for the purpose of measuring op-

Fig. 2. Customer records being checked for accuracy during conversion from the existing manual system to a new computer-based system.

erator performance with such a system. This is the mechanical simulation (MECHSIM) referred to by Lindgren in a 1966 article, in which he said: "The Bell project also presents an interesting sidelight. Human factors engineers always stress the importance of being in on design studies as early as possible so that their investigations can properly influence the final system design. In the Bell case, the engineering department is not even going to begin to implement the computer look-up system until they have seen the complete results of the human factors study . . . which amounts to the realization of a human factors dream" (7).

The dream, if that is what it is, has continued. The mechanical simulation of computer-assisted directory assistance, when it was completed in 1966, indicated that there probably would be sufficient savings in operator look-up time to justify the development of an actual computer-based system. As a result, a computer-based human factors experiment was begun in late 1970. It was referred to as the "live traffic experiment" because it involved ten operators who worked at specially designed CRT terminals and who received actual "live" customer requests for directory assistance. They used the terminals to locate customer telephone numbers which were stored in a computer memory.

Operators talked directly with customers and, using keys on the CRT console, keyed into the computer the necessary trigrams indicating name and address information (Fig. 1). The computer then activated a display on the CRT from which the operator selected the desired telephone number and relayed it to the customer.

The study was in two parts. The first was an experiment or, more correctly, the last stage of an experiment that had several preceding steps in other parts of the laboratories going back over some 12 years. Operators systematically tried eight search (keying) strategies to retrieve telephone numbers from computer storage. The best strategy was then selected for the second part of the experiment, a pilot study of the computer-based directory assistance operation. Its purpose was to see if sufficient savings in work time could be realized to justify the use of the computer subsystem for directory assistance.

The outcome of the pilot study indicated that the computer-based directory assistance would indeed save sufficient operator work time to justify a change to that type of system when development resources became available.

More important, perhaps, is the fact that the live traffic experiment represents in several ways the coming of age of human factors psychology or "human performance technology." For example, the human factors experiment and pilot study were undertaken *before* huge sums were committed to system development. Next, there was the specific commitment of designer resources to the attainment of an integrated operator-computer interaction. This required an explicit understanding that software designers were dependent upon human performance requirements developed by human factors psychologists.

Finally, the live traffic experiment served to call management attention to the principle that the personnel subsystem of a computer-based system requires careful design attention just as does the computer subsystem.

In information systems, the human-computer design is important for error prevention. Human-caused errors are an anathema to system designers. A large facilities engineering and assignment system that we studied illustrates the effect of human error on system performance; it also illustrates why, when systems have human-error problems, designers tend to focus their attention on interface and data display designs.

Almost immediately after the change to the new system, human errors snowballed. Computer-processed transactions bogged down to the point where the computer was bypassed entirely; later, the errors were located and corrected, and the computer data base and manual protective records updated. The system's error-correction process quickly became overburdened. System performance then fell below acceptable or even tolerable levels.

Most of the human errors were detected by computer edit of manual inputs. System designers proposed a two-part solution: Replace the input clerks with more skilled people; and substitute CRT terminals for input forms. Before this solution was implemented, a special task force worked with the system users. As a result, CRT's were not installed, nor were more skilled people hired. Instead, many of the paper forms were redesigned, performance aids developed, and system personnel trained. Error rates immediately dropped, and system performance was soon above acceptable levels, where it remained.

A key point, however, is this: The most visible changes made were to the forms used by humans to transmit information to the computer; however, task force members believed the other personnel subsystem work (procedures, performance aids, and training courses) to have been much more important.

This final example points out a critical issue. On many systems, the design of the man-computer interface itself is not

enough. After all, many large computer-based systems use off-the-shelf terminal equipment which satisfactorily meets human capabilities and stereotypes. As DeGreene says, in regard to design of data displays: "Many of the baffling human factors problems, then, relate not so much to vision as to cognition. The display can be considered a direct extension not only of the inner structure and workings of the computer, but also of what is going on within the user's [operator's] problem-solving head" (8).

For most system designers, human-machine interface design should focus on integrated, complementary human-machine procedures. Human procedures and machine procedures come together at the interface, whether console, CRT display, or paper form. The interface design should take into account these interactions, and should be consistent with both human and hardware/software capabilities.

Design the Subsystems Concurrently

Human performance technology integrates elements of human engineering, engineering psychology, instructional technology, and industrial and organizational psychology. The application of human performance technology is called personnel subsystem development (PSD), which means the integrated design of human procedures, human-machine interfaces, training, performance aids, and documentation as part of the total system. It also includes the rigorous testing of the personnel subsystem together with the computer and hardware subsystem counterparts. The products of PSD are the controlled utilization of human resources, and the means (procedures, training, performance aids, interfaces, for example) for obtaining required human performance in systems.

A recent Bell Laboratories example of designing the personnel subsystem simultaneously with the computer subsystem was the enhancement of a computer-based message switching system. During design there was continual interaction between personnel subsystem and software designers. The personnel subsystem and software were fully tested before installation. During a 3-month period following installation of the enhanced systems, only 6 percent of the problems encountered required design changes, and none of these involved human procedures or a human-machine interface. Previous experience suggested that, if integrated, concurrent design of the computer and personnel subsystem had not been done, 50 percent of the encountered problems would have required changes to the personnel subsystem (see 9).

Another example illustrates that even as a retrofit, the thorough design of the personnel subsystem of a computer-based information system can accrue system benefits. When the PSD design team began work, the system design included hardware and software. More than 50 documents already had been developed for system personnel by persons who did not follow the PSD process.

The PSD designers first analyzed system functions, then analyzed tasks to be done by system personnel. From the data collected, work procedures were designed. Common human procedures were identified. Performance aids were developed. As a result, the existing 50 procedural documents were reduced to 17. This in turn simplified document maintenance and reduced training time.

A key design activity was to allocate the human information load among system personnel and system products, without duplication. For example, attendants (operators) were taught to recall from memory certain procedures and data. Other data were contained in performance aids. Still other data were incorporated in masks displayed on the CRT. Attendants were trained to use the performance aids and CRT masks, not to remember all the information contained therein; there was no need to duplicate such information in training courses, therefore the volume of system documentation was reduced markedly.

This system also illustrates how manual procedures and interface design are interrelated. A service attendant at a CRT terminal talks directly with a customer who is reporting telephone malfunction. A mask is displayed on the CRT. As the customer gives information, the service attendant uses the CRT keyboard to fill in the mask. The original mask design had been controlled by software requirements, and assumed the typical customer would provide information in a certain order. Actually, the information from the customer came in a different order; therefore, attendants had to go through several extra operations to record the data, thus cutting productivity and increasing the chance of error (Fig. 2). Redesigning the mask to match the operator-customer dialogue reduced the number of operations required of the operator and reduced the chance of error.

One final example deals with the conversion of a system which maintains an inventory of available facilities, and assigns these facilities to customer orders for telephone service. The conversion requires the transcription of the records of 120 million telephones from paper to computer data base, at an estimated cost of $305 million.

The computer and personnel subsystems of the conversion system were designed and implemented concurrently and systematically. Training and documentation were developed, and the personnel subsystem was tested as part of the design process. When the testing program was completed, the transcription error rate had been reduced from a projected 10 percent to an overall 5 percent. Thus, when the necessary records have been converted to the computer data base, there will have been an estimated cost avoidance of $46 million.

Conclusion

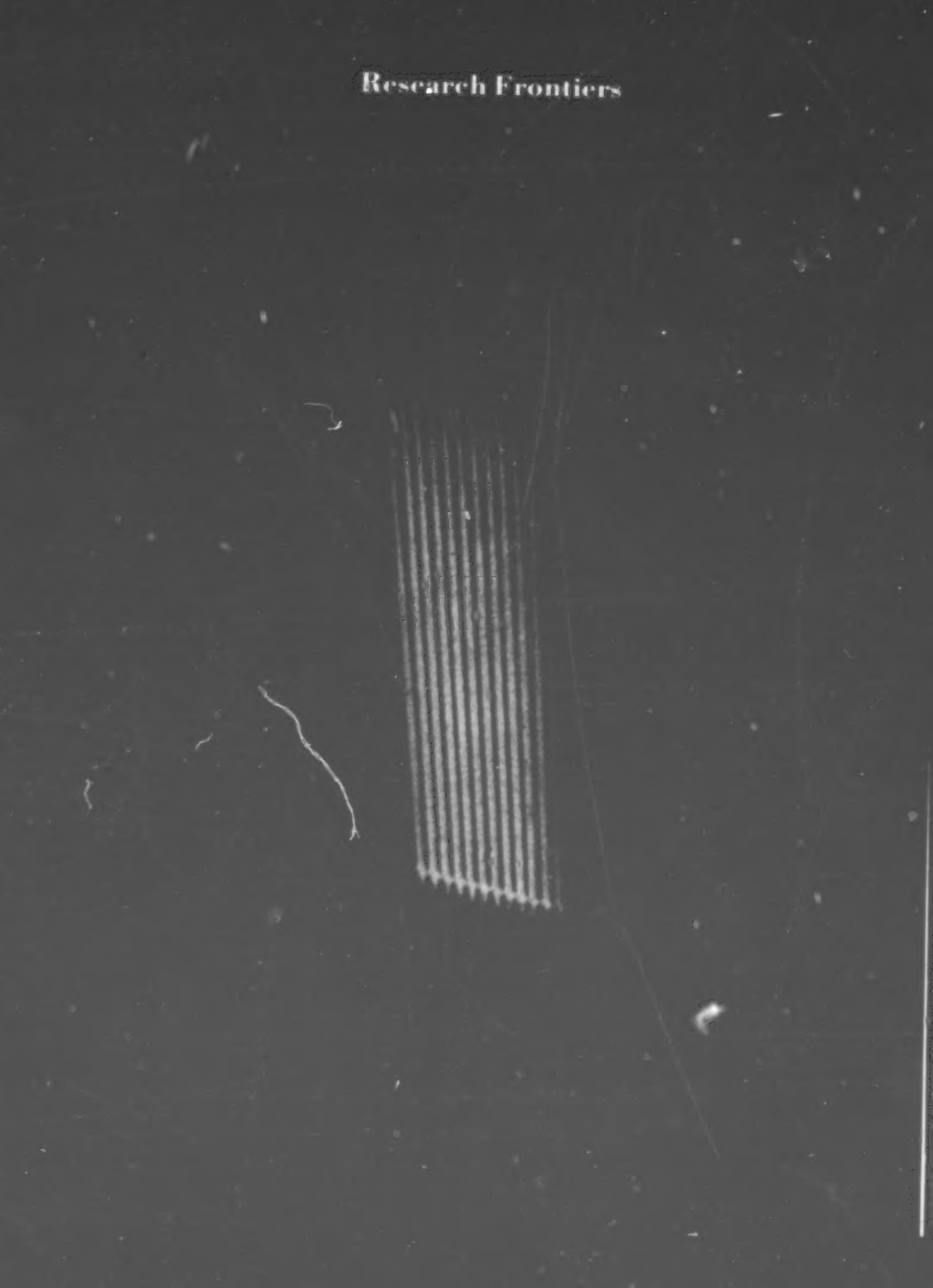
Clearly, the years since World War II have seen increasing complexity of human-machine interaction. Machines have become much more complicated and the responses required of humans may be said to have moved from "brawn to brains," that is, from the physical interaction between person and machine to an interaction which depends more on the information processing abilities of human beings.

In our experience with computer-based systems, we observe an increasing emphasis upon personnel subsystem development. This is largely because today's complicated systems, particularly those involving computers, put increasing demands upon people in the system for fast and nearly error-free performance. Careful design attention must be given to the human element if such performance is to be attained.

References and Notes

1. "Human Engineering in the National Defense," Panel on Human Engineering and Psychophysiology, Research and Development Board, HPS 205/1, (29 June 1953), cited by N. Lindgren, *IEEE Spectrum* **3**, (No. 3), 133 (1966).
2. H. P. Van Cott, *Human Factors Methods for System Design* (American Institutes for Research, Washington, D.C., 1960), foreword.
3. J. C. R. Licklider, *IRE Trans. Hum. Factors Electron.* **HFE-1**, PA-11 (March 1960).
4. V. A. Vyssotsky, private communication.
5. J. C. R. Licklider, *Comput. Autom.* **18**, 48 (August 1969).
6. W. F. Fox, paper presented at the Seventh International Symposium on Human Factors in Telecommunications, Montreal, Canada (1974).
7. N. Lindgren, *IEEE Spectrum* **3** (No. 4), 70 (April 1966).
8. K. B. DeGreene, *Systems Psychology* (McGraw-Hill, New York, 1970), p. 319.
9. M. W. Seth, "The Human Work Module: A Structural Entity for Personnel Subsystem Design" (Bell Telephone Laboratories, Murray Hill, N.J., 1976).

169



Photons in Fibers for Telecommunication

Glass fibers may replace copper wires for carrying voice, data, or video signals in telecommunications systems.

Stewart E. Miller

The prospective use of glass fibers in telecommunications is no accidental result or byproduct of another interest in electronics. Schawlow and Townes reported (*1*) the conception of a coherent oscillator or amplifier in the visible or near infrared region in 1958. It was based on stimulated emission and was related to a centimeter-wave device that was described 4 years earlier (*2*). The coherent visible source, which became known as the laser, was unique. No other highly coherent source existed in the visible region, and its potential usefulness in telecommunications was immediately recognized. The optical carrier (*3*), with a frequency 10^4 times higher than microwave carriers, could have superimposed on it an information bandwidth also larger by orders of magnitude.

Maiman built the first solid-state laser in 1960 (*4*), and Javan *et al.* (*5*) devised the gaseous laser in 1961. These and other developments had broad applications possibilities to communications, and led to rapid research on materials, optical devices, and optical transmission media. This research was reviewed by Kompfner in 1965 (*6*).

During the next 5 years a great deal of progress was made (*7*). Intercity communications proved technically feasible with a transmission medium consisting of periodically placed glass lenses, periscopic mirrors, or gas lenses, enclosed in an underground conduit system. However, large initial costs for such installations needed the support of large service requirements (over 10^{13} bit/second on an individual route) in order to make the system economics attractive. The glass-lens or periscopic-mirror beam waveguides were not adaptable to use through crowded conduits in the city. Either more refined gas-lens guides or an alternative was required.

Point-to-point radiolike transmission through the atmosphere was studied with the conclusion that reliable path lengths are limited to a few hundred meters by snow, fog, and rain. Such paths may become useful for special-purpose applications.

Glass fibers were seen as a conceptually attractive but unfortunately highly lossy transmission medium [typically over 1000 db (*8*) per kilometer]; the light-wave signal became "lost" in the noise after only a few meters of travel in existing fibers. An early publication by Kao and Davies (*9*) indicated very much lower losses in bulk fused silica, and this stimulated a concerted effort to make low-loss fibers by several groups. In Japan, workers proposed making a continuously focusing medium by grading the transverse index distribution in a glass fiber; they called it "Selfoc" and made a series of novel distributed image-focusing elements as well as a fiber transmission line (*10*). Theoretical work in this area came as early as 1965 (*11*), and a U.S. patent covering a graded-index fiber was issued in 1969 (*12*). In England, a consortium was formed to explore fibers for transmission, spurred by speculation that lower losses might be achievable. In all, four groups—Bell Laboratories, an English consortium, Nippon Sheet Glass–Nippon Electric, and Corning Glass Works—put forth strong efforts in parallel to try to bring losses down from values typical of fibers in medical instruments (1 db/m) to values of potential viability for telecommunications. The Corning group achieved significant results in 1970, producing a single-mode fiber hundreds of meters long in which light-wave loss is less than 20 db/km (*13*).

Later, MacChesney *et al.* (*14*) announced a new fabrication process, a modified chemical vapor deposition (MCVD) process which led to lower losses and improved fiber-design parameters. Numerous industrial and university groups are now making fibers in which losses are 2 db/km or less. The general size, physical flexibility, and cost of glass-fiber (fiberguide) cables resemble those same features of wire cables, and fiberguides are seen as potentially applicable wherever wire-pairs or coaxials are now employed. I shall show that light-emitting diodes (LED's) as well as lasers can be used as optical carrier generators; LED's are an attractive low-cost candidate for short-distance transmission in on-premises or aircraft or shipboard applications, while lasers make possible systems with much higher information rates such as might be used between cities or between central offices in the metropolitan area telephone plant.

Elementary Fiber Link

The optical portions of all such systems are similar and simple; they consist of a carrier generator (laser or LED), fiberguide cable, envelope detector [simple semiconductor-junction (PIN) or avalanching photodiode], and conventional electronics to drive the carrier generator and to follow the detector. The information rate or bandwidth required on the link, in combination with the desired link length, lead to the choice between laser or LED and to selection of one of the fiber types described below.

Two slightly different kinds of glass are used for the core and the cladding of fibers, giving the core a slightly higher index of refraction than the cladding. The electromagnetic-wave modes guided in the core have fields that decay radially in the cladding; with appropriate cladding thickness, very little influence on core modes is produced by the wave properties of the jacket or regions exterior to the jacket. Single-mode guidance is obtained with a core diameter of about 5 micrometers and an index difference between core and cladding of about 0.5 percent. Multimode fibers are necessary for carrying significant amounts of power from a LED; typical core diameters are in the 50- to 75-μm range with an index difference of 1 or 2 percent.

In the multimode fiber case there are subtleties associated with the transverse index distribution (Fig. 1) which have profound effect on the pulse-spreading or "delay distortion" property of the fibers. For so-called step-index fibers (uniform core index, $g \to \infty$ in Fig. 1), the

The author is director of guided wave research at Bell Telephone Laboratories, Crawford Hill Laboratory, Holmdel, New Jersey 07733.

—Courtesy Bell Laboratories

various core modes travel with markedly different group velocities. By grading the index the relative group velocities of the core modes can be made much more nearly equal, resulting in less pulse broadening. These effects can be stated in terms of digital transmission capability; for step-index fibers with $\Delta = 1$ percent (Fig. 1), the fiber capability is a bit-rate-distance product of 18 megabit-kilometers (Mb-km). This means that a fiber 1 km in length could carry digital signals at an 18-Mb rate without significant error, or alternatively, a 500-m-long fiber could carry signals at a 36-Mb rate. A near optimum grading of the transverse index distribution (a g value of about 2 in Fig. 1) raises this capability theoretically to around 14,000 Mb-km, or almost 1,000 times the capability of the step-index fiber (*15, 16*). Graded-index fibers having a transmission capability of 75 to 100 times the step-index fiber have been made in the laboratory, and research continues.

The largest transmission capability is theoretically available with single-mode fibers, in which wavelength dispersion proves to be limiting. Here with an 0.8-μm laser having a spectral width of 1 angstrom, the capability in fused silica is about 50,000 Mb-km.

The simplest and lowest expensive carrier generator for the fiber link is a LED; the message is imposed as a driving current variation, resulting in an emission brightness variation. The efficiency of coupling to the fiber is low, because of the incoherent nature of the emission; however, anywhere from 100 microwatts to a few milliwatts can be coupled into a telecommunications fiber, depending on the value of Δ (Fig. 1).

The semiconductor injection laser, in which stimulated electron-hole recombination emission is used, is a far more efficient carrier generator. With special care the laser can be made to oscillate in a single transverse mode, and most of the output can be coupled into the fiber. Continuous operation at outputs near 1 mw and pulsed low-duty-cycle output on the order of 1 watt have been achieved. At present, the materials used in such lasers are aluminum, gallium, and arsenic (AlGaAs system), with emission at 0.82 or 0.85 μm wavelength (*17*).

Attempts are being made to develop carrier generators with emission wavelengths of 1.0 to 1.5 μm, because (i) the fiber loss is lower at longer wavelengths (see Fig. 2), and (ii) fibers have less wavelength dispersion in the 1.0 to 1.5 μm region. Wavelength dispersion is still being studied, but it is known that the dispersion of pure fused silica has a

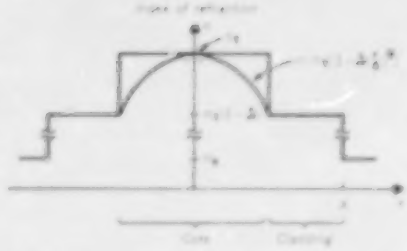


Fig. 1. Profiles of index versus radius for graded-index and step-index ($g = \infty$) fibers (fiber diameter = 2a). Symbols: a, fiber radius; r, distance from core center; g, parameter determining profile shape; n_c, surrounding refractive index).

minimum near 1.25 μm, and the dopants added move this optimal wavelength only slightly. At 0.82 μm the spectral width of the AlGaAs LED (about 350 Å) leads to pulse spreading in pure fused silica of about 4 nanoseconds per kilometer; this gives the LED-fiber combination a transmission capability of about 100 Mb-km as limited by wave-length dispersion. The AlGaAs injection laser typically has a spectral width of about 20 Å, leading to pulse spreading at 250 picoseconds per kilometer.

A new laser has been developed that has both longer emission wavelength (1.06 or 1.3 μm) and narrower spectral width (1 or 2 Å); this is the single-crystal Nd:YAG (neodymium doped yttrium aluminum garnet) fiber laser, pumped by an AlGaAs LED. Other contenders for future fiber systems use are the GaAsSb injection laser (*18*) and the InGaAsP injection laser (*19*), both of which have been operated as continuous wave oscillators.

The final optical element in the elementary fiber link is the detector. Either the

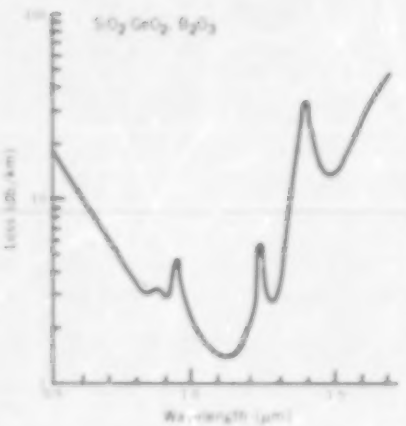


Fig. 2. Loss plotted against wavelength for a good multimode silica fiber in which GeO_2 and B_2O_3 were used as dopants. [From Kaiser and MacChesney (*17*)].

simple PIN silicon diode or the avalanching photodiode is employed; the hole-electron pairs created by the incident photons are swept out by a static applied field, yielding a conduction current carrying the modulation of the incident light intensity. The PIN is less expensive, but yields a weak signal which may be obscured by thermal noise in the following amplifiers. The avalanching detector yields a current larger by 10 to 100 times, and this rides above the thermal noise of following amplifiers. The avalanching detector is more complex and therefore more expensive, and also requires a precisely set bias potential which must be reset to follow ambient temperature changes. The net result is to reserve the higher-cost high-performance avalanching detector for those applications where the performance of the PIN detector is inadequate (or where the incremental cost of the avalanching detector assembly results in system savings elsewhere).

Silicon detectors are useful at wavelengths up to about 1.06 μm; beyond that the absorption falls off catastrophically and either germanium or one of the semiconductor compounds must be used. Only silicon detectors meet all performance requirements now (such as low dark current and reproduceability), but with the interest in the 1.0- to 1.5-μm region already noted investigators are looking for other satisfactory materials.

Fiber and Cable Fabrication

One of the most novel aspects of fiber-guide transmission is the science and technology of telecommunications fiber and cable fabrication. But let us consider first the carry-over of older techniques. The rather remarkable formability of glass is utilized in a very straightforward way (Fig. 3), which also shows enlarged view. The preform is on the order of 1 or 2 cm in diameter and 40 to 100 cm long, and is a scaled-up model of the fiber desired. The center of the preform is made of the glass composition desired in the fiber core, usually nearly pure silica with a few percent of an additive, or dopant, such as germanium, boron, phosphorus, or combinations of these elements. The outer shell of the preform is made of the glass desired as the fiber cladding, frequently silica with boron as the dopant. Then in a single operation the fiber is drawn, a 100 to 1 diameter-reduction ratio and a fiber-pulling velocity of 1 m per second being typical. New techniques for controlling diameter (on-line measurement plus feedback) and for preserving high strength have been add-

ed to the system as a consequence of the potential telecommunications applications of these fibers.

The preform needed is a recent development. The preferred graded-index distribution of Fig. 1 required a new approach, as did the need for glass purity of a new order. There are of course two broad classes of losses to be minimized: absorption loss, caused in part by electronic transitions in impurity ions, and scattering losses caused by inhomogeneities. Figure 4 shows the added bulk losses to be expected in pure fused silica with the impurities noted. In order to get down near the intrinsic Rayleigh scattering limit one needs no more than one part in 10^8 of the OH radical, three parts in 10^9 of Cu^{2+}, or one part in 10^9 of Fe^{2+}, for example. Other transition elements are similarly troublesome. Systems containing high concentrations of silica have proved most practical. Two methods of fabricating high-silica preforms have appeared; the first is the soot-deposition process (*20*) and the second is the modified chemical vapor deposition (MCVD) process. Both methods have produced fibers in which light-wave losses are very low, although the MCVD process appears to have been more widely adopted because of its simplicity and because it can inject higher percentages of dopants. Adjustment of the doping level of the successive layers of glass according to a programmed plan can yield the transverse index profile desired. In Fig. 2, the fiber doped with Ge and B shows peaks near 0.95 μm, 1.2 μm, and 1.4 μm which are all due to the presence of OH^-. Even greater purity has been reported (*21*) by several Japanese workers, with material losses dipping below 1 db/km near 1.2 μm.

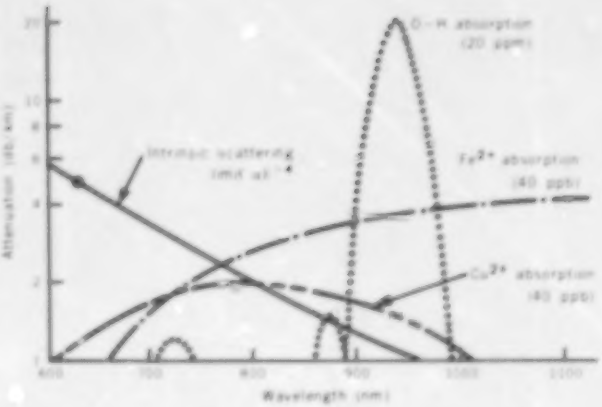


Fig. 4. Typical loss contributions for fused silica. Concentrations of impurity ions for O–H, Fe^{2+} and Cu^{2+} lead to the absorptions shown. For pure fused silica the intrinsic Rayleigh scattering loss, varying as λ^{-4}, is also shown.

However, a low-loss fiber is just the beginning of achieving a low-loss fiberguide cable. The most novel aspect of cabling optical fibers stems from an electromagnetic wave propagation effect: axial kinks, or minute changes in direction at length intervals of the order of 1 mm, such as can be created by winding a fiber tightly on a drum having a rough surface, can cause coupling from the core-guided modes to a radiation field and can easily result in added losses in the 5 to 50 db/km range (*22*). A solution, proposed by Gloge (*23*), is to form a minute composite structure which isolates the fiber from external mechanical influences. However it is done, the fiberguide must be constructed in such a way as to avoid "microbending" losses, as they are called.

In the most common form of cable, the communications fibers are protected by plastic jackets and strengthening components, and the communications fibers are handled individually at the cable ends. Such cables have been made at Bell Laboratories, where another approach has also been explored. The fibers are first grouped into ribbon-like arrays (*24*) so that all subsequent splicing and handling can be done on the ribbon assembly. Figure 5 shows a ribbon assembly (12 fibers per ribbon, 12 ribbons per cable) in end view and in a completed splice. Here the 144 fibers were first aligned in grooved chips (preferentially etched silicon chips) and faced off to form a solid array which was mass spliced to an identical mate (*25*). A 658-m section of this cable was used in a fiberguide system experiment at Atlanta. In that section, 138 of the 144 fibers survived the cable-making process; no fiber breakage occurred during installation in ducts reasonably typical of commercial installations; and the final mean loss of all surviving fibers was 6.0 db/km at 0.82 μm, an increase of 1.3 db/km over the mean uncabled loss of the same fibers, the change being attributed to microbending (*25*).

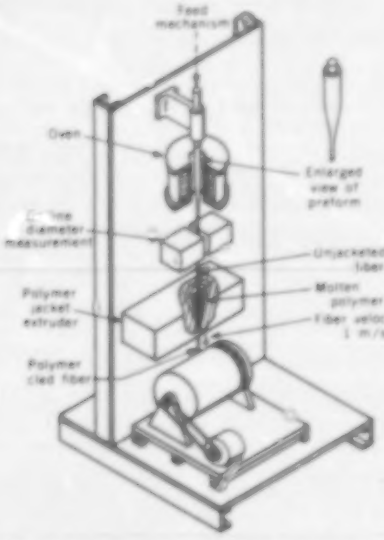


Fig. 3. Diagram of a telecommunications fiber-drawing bench.

Fiber Connectors and Splicing

More new structures are required in fiberguide transmission systems to provide both permanent splices and fiberguide connections which can be disengaged. The background of theory can be summarized as follows: The requirement on transverse alignment of the fiber ends is about 0.2 core radius for approximately 0.2 db loss. For single-mode fibers, 0.2 core radius is about 1 μm, a requirement that is likely to be expensive to meet. For that reason, single-mode fibers are likely to be used only where the single-mode performance is essential.

Longitudinal end separation can be on the order of a core radius for about 0.1 db loss (with a bridging fluid or solid at the joint), so this requirement is not as critical as that for the transverse align-

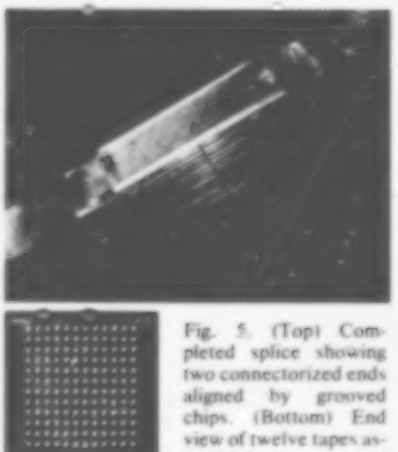


Fig. 5. (Top) Completed splice showing two connectorized ends aligned by grooved chips. (Bottom) End view of twelve tapes assembled in grooved chips for use as a prefabricated connector (*38*). [Courtesy of M. I. Schwartz]

ment, especially for multimode fibers where the core radius is about 25 μm.

In a splice or demountable connector it is essential to have a properly prepared fiber end. A refinement of the "score-and-break" technique, in which the fiber is subjected to simultaneous tension and controlled-radius bending, yields ends that are smooth and perpendicular to the axis. This has been demonstrated not only on single fibers (26), but also on tape arrays of fibers (27) wherein the entire array is prepared in a single step.

Single-fiber splices have been accomplished by several groups. Such splicings were achieved by slipping the two ends into a close-fitting round sleeve and gluing them together as a butt joint. This yielded rather high losses when fiber diameter tolerances were accommodated. In an innovation of this method, the ends are assembled within a square tube in the presence of a bending moment. The fibers align themselves accurately in one of the corners, and losses on the order of 0.1 db are reproducibly achieved (provided that the fiber diameters are comparable).

Figure 6 shows a single fiber connector that has found application not only in "patch cord" applications but also as a plug-in fiber connector on plug-in electronic circuit assemblies. A molded male with tapered end mates with a biconical female; a compliant solid bridges the gap across the fiber ends. Mean connection losses under 0.4 db have been achieved (28).

The splicing of an entire tape array of fibers has been accomplished in a single step by aligning the fibers in an embossed chip and sealing them with epoxy glue (29).

Transmitter-Receiver Electronics

In a fiber-optic telecommunication system, the electronic circuits which drive the laser-LED or which follow the detector are mostly conventional, but there are a few critical points to note. The laser has a sharp threshold of oscillation, and must not be driven very far above that point or damage may result. Furthermore, the threshold current changes during the life time of the laser and is sensitive to ambient temperature changes. As a result, stabilizing electronics is required to control the laser's operating point. Similarly, the avalanching photodiode detector (APD) requires a static bias critically related to avalanching gain and signal-to-noise ratio; this bias must also be changed as the ambient temperature changes. The solution is another auxiliary circuit. For the above reasons the receiver electronics is less expensive for LED-transmitter and simple photodiode-receiver arrangements than for the laser-APD arrangement; the latter is likely to be used only where the larger signal-to-noise range is required.

Initial fabrication of repeater assemblies has combined the current technology, in which plug-in circuit boards and thin-film or integrated circuit assemblies are used, with the optical elements. Further integration of the fiber connector with the associated electronics will probably be achieved in the future, so that the result will be a single package (24).

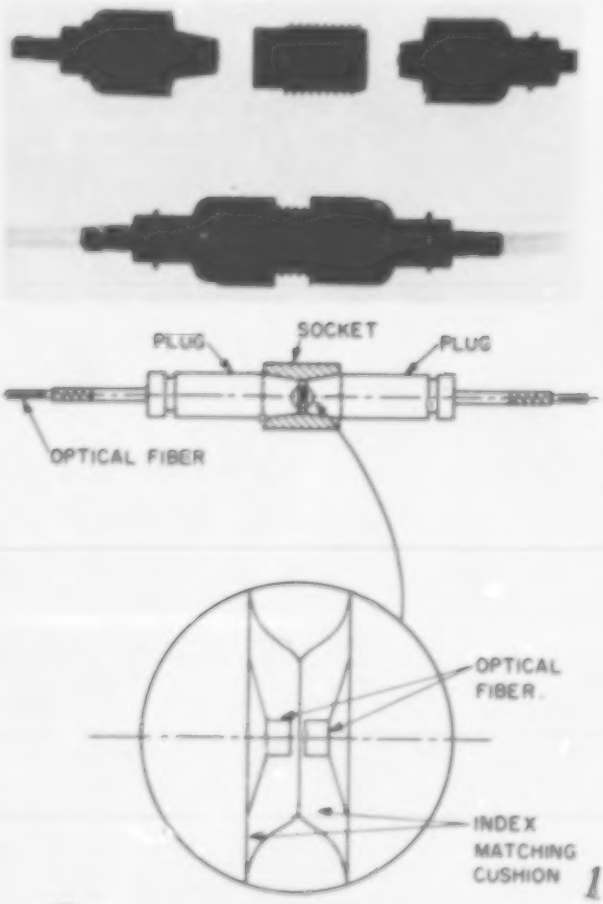


Fig. 6. Single-fiber connector designed by P. K. Runge of Bell Laboratories. The two identical males mate with a biconical female and yield a mean connection loss under 0.4 db. Fabrication of both males and female is by transfer molding, an inexpensive process (28).

Modulation Format

The modulation format is usually dictated by the signal being transmitted, and either analog or digital transmission can be used. It is noteworthy that the photodiode detector yields electrons for photons received, so the output modulation current is directly proportional to lightwave power input—a square-law relationship. This works out to favor a signal format with a large peak factor for favorable ratio of signal to thermal noise.

For digital signals, Fig. 7 shows the required number of photoelectrons per bit (or received signal power level) as a function of bit rate (15, 30). The upper band implies simple photodiode (PIN) detectors, and the lower band implies APD detectors with optimal avalanche gain ranging from 25 to 100 as indicated on Fig. 7. Research model receivers performing within these bands have been built with both PIN and APD detectors at 6.3 Mb/second, 45 Mb/second, and 274 Mb/second (15).

Digital transmission can be advantageous, not only where the customer's signal is digital (as with computers or data processing), but also where an analog (continuously varying) signal such as voice or video is to be transmitted. The analog signal is periodically sampled, and each sample value sent as a coded group of on-off pulses. Hence the name pulse-code modulation (PCM). A major advantage is that the PCM signal can be regenerated to eliminate accumulated noise at each amplification point, permitting far longer systems than is feasible with analog transmission. The disadvantage of PCM is the cost of the coder-decoder which is required.

A preferred form of analog transmission for lightwave systems is pulse-posi-

174

tion modulation (PPM); here the spacing between pulses is continuously modulated by the analog signal, thus eliminating the need for the coder-decoder. The PPM system has been shown capable of transmitting a voice channel at 70 db peak ratio of signal to r.m.s. noise ratio (where r.m.s. is root mean square) through a fiberguide loss of 70 db with a LED transmitter and PIN detector (*31*), or more than about 100 db fiber loss with a pulsed laser. At today's low fiber losses, this implies that very long fiber spans might be used.

In the system noted above, PPM uses expanded bandwidth on the fiber to gain a signal-to-noise advantage, and since fibers have broad bandwidths, this is advantageous. Waveform sampling is required for PPM, with some associated complexity. An alternative way of using bandwidth expansion and avoiding the linearity requirements of simple intensity modulation is to use a frequency-modulated (f-m) subcarrier to modulate the LED or laser. This technique should be useful, particularly where an f-m analog signal may be necessary or convenient for other reasons.

Comparison with Wire-Pairs or Coaxial Cables

Lightwave transmission on fiberguide has several advantages over other means of telecommunication. (i) Small size (smaller and lighter than either wire pairs or coaxials). (ii) The fibers are tolerant of bends in the centimeter range. (iii) The fibers are nonconducting and noninductive. This is valuable in avoiding spurious ground currents, the interference effects of lightning, and possibly also the damage effect of lightning. The fiberguides do not radiate in the usual radio spectrum, that is, they are not generators of interference, nor do they pick it up. (iv) The fibers have lower losses than wire pairs or coaxials, at least for bandwidths above a few megacycles. (v) Much wider bandwidths are available with these fibers than with wire pairs at (prospectively) modest extra cost. (vi) Negligible signal interaction (or cross talk) occurs between adjacent fibers in the same cable.

Cost, the final element of comparison, is all important but difficult to pin down. Cost is extremely sensitive to volume of production. Because the principal components of fibers are abundant, in the long range at large-volume production they should become inexpensive. They are now produced in small volume and are quite expensive (approximately $10 per meter for a six-fiber cable).

Short Length Applications

There are numerous applications for short lengths of these fibers under consideration. In mid-1973, an experimental 16-Mb 30-m fiberguide link was installed in a laboratory model digital switching machine (No. 4 ESS) with a Burrus LED, a plastic-clad fused-silica fiber made at BTL Crawford Hill, and PIN receiver. Potential for satisfactory performance was established and an exploratory development is now under way (*32*). There are numerous 30-m to 400-m coaxial-cable links in the present design No. 4 ESS machine; the fiberguide alternative would be smaller, lighter, and free of interference pickup or generation. For military applications similar short links have been used on aircraft and on board ships (*33*). The use of fiberguide links in computing machines and in commercial measuring apparatus is also being considered.

For all of these links it is anticipated that LED-PIN technology is most appropriate, and once they are produced in volume these elements will be inexpensive. For some of these applications, the lifetime of the device may not be too critical, but in the No. 4 ESS machine, for example, the lifetime of the fiber links should be in the region of 10^5 to 10^6 hours to keep failure rate within system requirements. Reliability of this type is generally difficult to achieve with an infant technology.

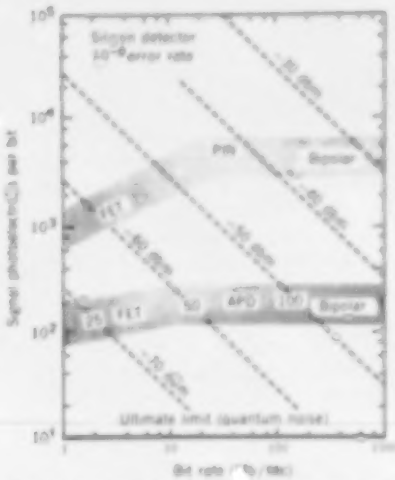


Fig. 7. Theoretical prediction of the required photons per bit for binary digital transmission at the abscissa rate with an error incidence of 10^{-9}. The dashed lines show the corresponding power level of the received signal. The band labeled PIN corresponds to use of PIN detector and low-noise electronics. The band labeled APD corresponds to use of an avalanching photodiode detector. The labels FET (field effect transistor) and "bipolar" refer to the preferred type of transistor amplifier following the detector. [Chart conceived by T. Li, based on theory of Personick (*30*)]

Fiberguide System Experiment at Atlanta

As part of Bell Laboratories' broad research effort on optical communications and to provide a focus for certain key specific development activities, planning for a fiberguide system experiment began in 1974 (*34*). Of the several prospective Bell System uses, the one selected for this exploratory effort was transmission between switching centers. Currently, extensive use is made of a pulse code modulation system which transmits 24 voice channels on a single pair of wires. The lightwave system alternative could employ any one of a number of channel groupings on a single fiber, depending on service requirements and economic tradeoffs. The Atlanta experiment was carried out initially with a 672-channel format and a 44.7 Mb/sec pulse rate.

A 144-fiber cable was fabricated in a 1-km length and cut down to fit into an existing standard underground plastic duct (*25*). Studies of fiber loss and bandwidth were made on the fibers before cabling, after cabling, and after installation in the duct. This experience serves as a foundation for any future cable design and, indeed, the initial performance fully met the designers' expectations. Further research and development will be beneficial in the areas of index profile control, structures for minimizing cabling loss, and processes for maximizing the strength of individual fibers (*35*).

The Atlanta experiment included electronics to control the avalanche photodiode and amplify its output, to perform the digital decision and retiming function, and to drive the injection laser. By using optical patch cords (short sections of fiber) a new fiberguide distributing frame containing 288 connectors, each associated with a fiber end, permits one to connect any of the various fibers of the 144-fiber cable to any of the repeater assemblies on an adjacent bay. Some of the best fibers were patched in series with 11 repeater assemblies to form a fiberguide path 64 km long with error-free transmission. The results from the Atlanta experiment have been extremely encouraging (*32, 34*).

Other Applications

There are other prospective applications for fiberguide. One is in the local loop, the connection between the customer and the central office. Increased bandwidth, at much less increased cost than would be necessary if wires were used, is anticipated.

In intercity installations, high-capacity

fibers may some day find a place; the biggest questions are probably economic rather than technologic, although research is needed to clarify which of the alternative pathways to attaining high capacity is to be preferred.

Fiberguide might also be used to provide control and communication for electrical power companies, where the insulating properties and noninductive nature of fiberguide would be great assets (36).

Although the United States has held a leading position in this technology, much research into the use of fibers for telecommunications is being conducted in Japan, where the service to power companies has received great emphasis (36). There is also considerable interest in the art in England, Germany, and France.

The Future

The pace of activity is feverish, expectations are high, and economics continues to regulate the ultimate usage of glass fibers in telecommunications. With so many fiberguide potentialities and so much effort to find economically viable applications, photons in fibers seem certain to flourish.

References and Notes

1. A. L. Schawlow and C. H. Townes, *Phys. Rev.* **112**, 1940 (1958).
2. J. P. Gordon, H. J. Zeiger, C. H. Townes, *ibid.* **95**, 282 (1954).
3. In accordance with radio terminology, the term "carrier" here refers to an electromagnetic wave or current, usually sinusoidal, which is modulated by the information signal to be transmitted.
4. T. H. Maiman, *Nature (London)* **187**, 493 (1960).
5. A. Javan, W. R. Bennett, D. R. Herriott, *Phys. Rev. Lett.* **6**, 106 (1961).
6. R. Kompfner, *Science* **180**, 149 (1965).
7. S. E. Miller, *ibid.* **170**, 685 (1970).
8. The rate of power loss as a wave travels in the direction of propagation is defined as

$$10 \log_{10} \frac{\text{input power}}{\text{output power}}$$

Power loss in decibels =

The maximum tolerable power loss between amplifiers in a communication system is usually about 50 db, or a factor of 10^5.
9. K. C. Kao and T. W. Davies, *J. Sci. Instrum.* **1** (Ser. 2), 1063 (1968).
10. T. Uchida, M. Furukawa, I. Kitano, K. Koizumi, H. Matsumura, *IEEE J. Quantum Electron. QE-5* (Abstr.), 331 (1969).
11. S. E. Miller, *Bell System Technical Journal*, **44**, 2017 (1965).
12. S. E. Miller, U.S. Patent 3,436,774 (25 March 1969).
13. F. P. Kapron, D. B. Keck, R. D. Maurer, *Appl. Phys. Lett.* **17**, 423 (1970).
14. J. B. MacChesney, P. B. O'Connor, H. M. Presby, *Proc. IEEE* **62**, 1280 (1974).
15. S. E. Miller, E. A. J. Marcatili, T. Li, *ibid.* **61**, 1703 (1973).
16. E. A. J. Marcatili, *Technical Digest*, OSA-IEEE Topical Meeting on Optical Fiber Transmission, Williamsburg, Va., 22–24 February 1977.
17. W. B. Joyce, R. W. Dixon, R. L. Hartman, *Appl. Phys. Lett.* **28**, 684 (1976).
18. R. E. Nahory, M. A. Pollack, E. D. Beebe, J. C. DeWinter, R. W. Dixon, *Appl. Phys. Lett.* **28**, 19 (1976).
19. J. J. Hsieh, J. A. Rossi, J. P. Donnelly, *ibid.*, p. 709.
20. R. D. Maurer, *Proc. IEEE* **61**, 452 (1973).
21. H. Osanai, C. Shioda, T. Moriyama, S. Araki, *Electron. Lett.* **12**, 549 (October 1976).
22. W. B. Gardner, *Bell System Technical Journal*, **54**, 457 (1975). R. D. Maurer and his associates at Corning Glass Works called this loss to our attention at an early date.
23. D. Gloge, *ibid.*, p. 245.
24. R. D. Standley, *ibid.* **53**, 1093 (1974).
25. M. J. Schwartz, R. A. Kempf, W. B. Gardner, *Digest of Second European Conference on Optical Fiber Communication, Paris, France* (September 1976), pp. 311–314.
26. D. Gloge *et al.*, *Bell System Technical Journal*, **52**, 1579 (1973).
27. C. M. Miller, *ibid.* **54**, 1215 (1975).
28. P. K. Runge, I. Curtis, W. C. Young, *Technical Digest*, OSA-IEEE Topical Meeting on Optical Fiber Transmission, Williamsburg, Va., 22–24 February 1977.
29. E. L. Chinnock, D. Gloge, P. W. Smith, D. L. Bisbee, *Bell System Technical Journal*, **54**, 471 (1975).
30. S. D. Personick, *ibid.* **52**, 843 (1973).
31. W. S. Holden, *ibid.* **54**, 285 (1975).
32. W. M. Hackett, J. R. Jones, C. A. Brackett, L. C. Dombrowski, L. E. Howarth, P. W. Shumate, R. G. Smith, A. W. Warner, R. S. Riggs, *Technical Digest*, OSA-IEEE Topical Meeting on Optical Fiber Transmission, Williamsburg, Va., 22–24 February 1977.
33. L. Dworkin, K. Trumble, D. Williams, *Technical Digest*, OSA-IEEE Topical Meeting on Optical Fiber Transmission, Williamsburg, Va., 22–24 February 1977.
34. I. Jacobs, *Bell Lab. Rec.* **54**, 291 (1976).
35. S. Matsuoka, *Technical Digest*, OSA-IEEE Topical Meeting on Optical Fiber Transmission, Williamsburg, Va., 22–24 February 1977.
36. M. Nishida, K. Fukatsu, T. Asahara, Y. Ueno, M. Shimizu, T. Yasuo, *Technical Digest*, OSA-IEEE Topical Meeting on Optical Fiber Transmission, Williamsburg, Va., 22–24 February 1977.
37. P. Kaiser and J. B. MacChesney, unpublished data.
38. C. M. Miller, *Bell System Technical Journal*, **54**, 1547 (1975).

New Technologies for Signal Processing

Charge-transfer and surface acoustic-wave phenomena are used in advanced signal processing devices.

Robert W. Brodersen and Richard M. White

The signal processors discussed here—the charge-transfer device (CTD) and the surface acoustic-wave (SAW) device—are based on new electronic design principles. Charge-transfer and SAW devices use an analog representation of information which moves along a path on a planar surface: in the CTD charge packets move in a semiconductor, and in the SAW device elastic wave energy propagates. Short-term storage of charge is possible in the CTD, and the speed of motion of charge through the device can be controlled externally, providing the opportunity to accept information at one rate and output it at another. In both the CTD and the SAW, simple means exist for sensing the moving charge or wave frequently along its path; with either technology this permits one to realize new types of signal processors based on the concept of the transversal filter, which is discussed below.

Charge-transfer and SAW devices are in similar stages of evolution. Their development until recently has been primarily in research laboratories which were perfecting the technologies. Both technologies are now sufficiently well understood that designers are beginning to incorporate them in complex electronic systems. A few of these new system applications and the advantages obtained by using either a CTD or an SAW are presented in this article. We expect that an accelerating use of these devices will be seen as system designers become more familiar with them and the devices themselves become more readily available.

Robert Brodersen is an assistant professor and Richard White is a professor in the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley 94720.

Charge-Transfer Device Delay Lines

A CTD is an array of closely spaced capacitors fabricated by using metal oxide semiconductor (MOS) technology. Its operation involves the movement of a charge packet which is stored on one

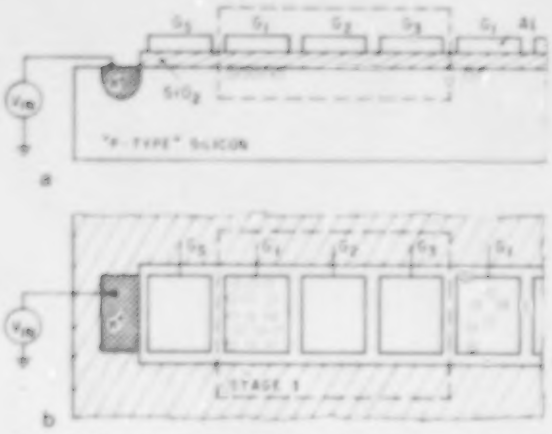


Fig. 1 (left). (a) Cross section of a portion of a CTD with an input composed of an n^+ diode and sampling gate G_1 which isolates the signal charge packet. The first stage of the device is enclosed by dashed lines. (b) Top view corresponding to (a).

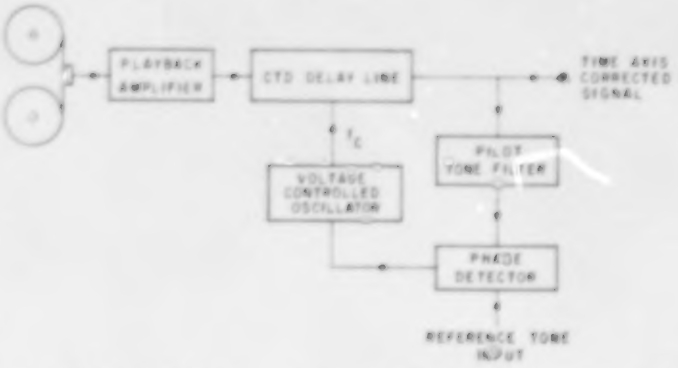


Fig. 2 (right). Timing error correction system for tape recorders, using the variable delay capabilities of a CTD.

capacitor to an adjacent capacitor in the array. In addition to signal processing, CTD's are finding important use as digital memories as well as solid-state imagers (*1*). The simplest application of a CTD is as an electronically variable analog time delay. In this section we briefly discuss the basic operation of a CTD (*2*) delay line and two applications that are made possible by this unique CTD capability.

A cross section of a typical CTD structure is shown in Fig. 1a. The device is composed of a silicon dioxide insulating layer 1000 angstroms thick sandwiched between a metal layer (typically aluminum) and a semiconducting (silicon) substrate, which for CTD's is generally doped to be *p*-type. A CTD is formed by selective etching of the metal (see Fig. 1b), localized oxidations, and diffusion of elements such as boron and phosphorus into the silicon. Its features can be extremely small (on the order of micrometers), and photographic techniques are used to define the circuit patterns.

Application of a positive voltage to one of the metal gates depletes the *p*-type silicon underneath that gate of holes, which are its major charge carriers. However, this voltage attracts the minority carrier electrons, because it produces an energy minimum (potential well) for electrons at the interface between the silicon and the insulator. It is these potential wells that are used for storing and transferring the signal charge.

The potential wells can fill up with electrons much as a container can fill up with fluid. There are two sources of these electrons: a controlled number of electrons (a signal charge packet) can be introduced into a well formed under the first gate G_1 by means of an external input voltage, $V_{in}(t)$, or electrons can be randomly introduced throughout the CTD by thermal generation from within the silicon. The CTD's are therefore dynamic devices in which the desired signal charge can be stored only for intervals that are short enough so that the thermal generation is sufficiently small. This time is typically of the order of seconds at room temperature.

To transfer the charge from one MOS gate to another, appropriate clocking pulses (typically 10 to 15 volts) are applied to the metal gates. To transfer the charge from gate G_1 to the adjacent gate G_2, a positive voltage is applied to G_2 (creating a potential well beneath it) while the well under G_1 is collapsed by reducing the G_1 voltage. The electrons that were stored in the G_1 potential well will now transfer along the semiconductor-insulator interface to the new well under G_2. A similar transfer moves the charge from G_2 to G_3 and then from G_3 to the next G_1 gate outside the dashed lines in Fig. 1.

The three gates enclosed by the dashed lines in Fig. 1 are called a stage; one complete clock cycle of T_c seconds is required to move the charge packet one stage to the right. When the charge packet transfers into the second stage it is sufficiently isolated that another packet can be introduced into the first stage without causing mixing of the two packets. In this way N charge packets, which are analog representations of N time samples of $V_{in}(t)$ (one sample every T_c seconds), can be stored simultaneously in an N-stage CTD. Since the delay through one stage is T_c seconds, N stages yield a delay of NT_c seconds. Therefore to change the time delay it is only necessary to change the period, T_c, of the external clocking waveforms. It is important to note that even though the input signal is sampled in time (as in the case of digital processing) the amplitude of the signal is retained in analog form. This makes it possible for the CTD to have the stability and convenience of operating under clocked timing while retaining the computational efficiency of an analog approach (an advantage that will be shown later).

To sense the delayed input signal, it is necessary to convert the charge packet back to either a voltage or a current. A voltage can be generated by transferring the charge packet onto an output capacitance, which has a buffer amplifier attached to it.

An important characteristic of CTD circuits is their high density, which stems from the use of the MOS technology as well as their extremely simple repetitive structure. A single stage of a CTD designed for signal processing typically measures 25 by 250 micrometers, which means that a 100-stage device is only 0.25 by 0.025 centimeter. Since a typical MOS integrated circuit measures 0.5 by 0.5 cm, this CTD requires only about 2 percent of the total available circuit area. It is therefore desirable in systems applications to integrate other system functions onto the same integrated circuit (IC) as the CTD. This capability of system integration is expected to provide much of the impetus for future developments in CTD signal processing.

Two applications of the CTD delay line that make use of the ability to vary the transfer rate of the charge packets (by varying the clocking period T_c) are transient data recording (*3*) and the correction of timing errors in tape recorders (*4*).

Many physical phenomena occur so quickly that it is very difficult to analyze

the data from a single event as it happens. It is therefore desirable to store the data and then perform the analysis at a more moderate rate. Unfortunately, most conventional storage techniques require that the data be converted first into a digital representation with a high-speed analog-to-digital converter, which can become extremely expensive for short sampling intervals. A CTD provides a very inexpensive and convenient means of sampling and storing a data transient in analog form. Since the CTD sample period can be varied, after a number of samples have been taken at sufficiently short sample intervals (that is, a high enough clock rate $1/T_c$), the clock rate is reduced and the data are read out at a slower rate for analysis. Devices that can store 1000 samples taken at time intervals as short as 50 nanoseconds (a clock rate of 20 megahertz) are commercially available, and clock rates as high as 130 Mhz have been achieved with special CTD designs (5).

Because the clock rate of a CTD is variable, it is also possible to have an electronically variable time delay, since the total delay time through an N-stage delay line is NT_c. This ability is very useful for correcting small, rapidly varying timing errors such as those which arise in video tape recorders because of an out-of-round capstan, fluctuations in the drive motor speed, and tape stretching. (Long-term timing errors can be corrected by the slower conventional motor speed control techniques.) Figure 2 shows a CTD timing error correction system (4) in which a steady pilot tone is recorded on the tape along with the desired signal. During playback the signal and pilot tone are delayed by a CTD delay line. The delayed pilot tone is then compared to a reference tone of the same

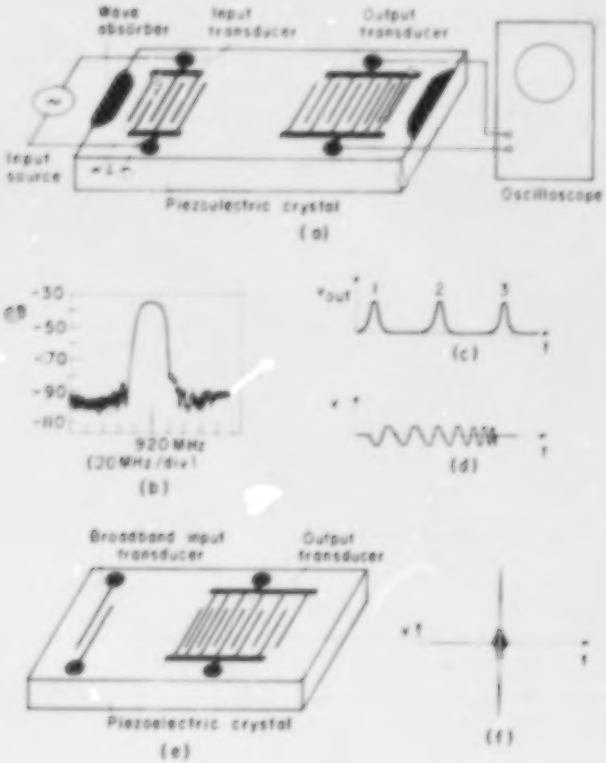


Fig. 3 (left). Operation of SAW electrode transducers. (a) Schematic of a typical SAW device, showing the voltage source connected to the input transducer producing surface waves that propagate to the output transducer, which has nonuniformly spaced fingers. The output voltage might be displayed on an oscilloscope as shown. The strongest wave excitation by the input transducer occurs if the distance L equals one SAW wavelength at the input source frequency. (b) Bandpass filter characteristic (insertion loss versus frequency) resulting from frequency-dependent excitation and reception by transducers in an actual SAW filter (21). (c) Frequency analyzer application showing the output voltage V plotted against time delay t in a device like that in (a) when low, medium, and high frequencies are input (outputs 1, 2, and 3, respectively). (d) Chirp output that would be obtained from the device in (a) if a voltage pulse of very short duration were applied to a broadband input transducer consisting of a single pair of electrodes. (e) Sketch of an SAW device having a broadband input transducer and an output transducer identical with the device in (a) but reversed right to left. (f) Sketch of the output when the expanded chirp pulse (d) is applied to the input of the device in (e). [The voltage and time scales are the same as in (d).] Note that the pulse has been compressed in time and increased in amplitude, permitting detection and precise timing in a noisy environment.

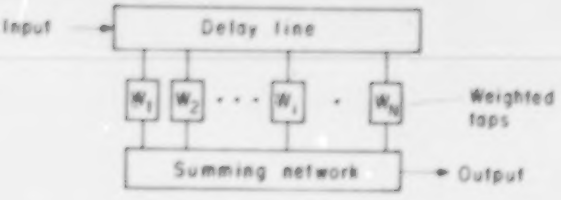
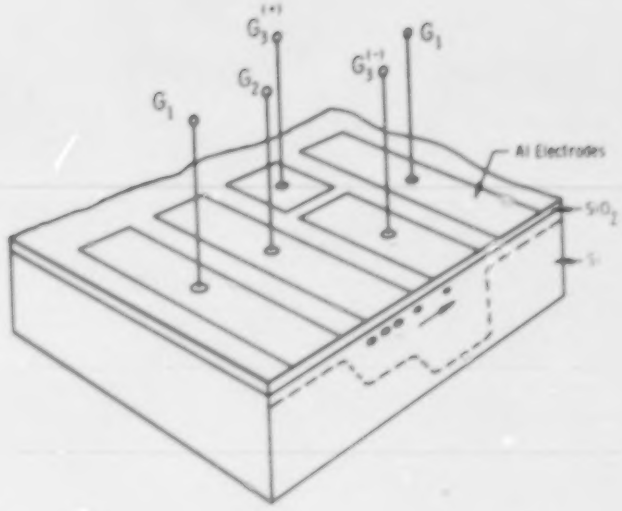


Fig. 4 (above). Schematic of the transversal filter, which is the basis for the design of many CTD and SAW devices. Fig. 5 (right). Portion of a CTD transversal filter, showing the split in one electrode (G_3) that determines the weighting coefficient.



frequency, and timing variations can be detected as a phase change between the recorded pilot tone and the reference tone. The phase detector shown in Fig. 2 will produce a voltage proportional to this phase difference which is used to control the frequency of an oscillator in a standard voltage-controlled oscillator circuit. This frequency determines the clock rate (and thus the delay) through the CTD in such a way as to eliminate the timing error that was detected. The advantage of such a system is that it could be integrated on a single IC and thus could provide a very inexpensive technique for improving recorder performance.

There are many other applications of analog time delay (6), including echo generation for electronic music, ghost cancellation in television, and filtering. Also, by the addition of input (or output) structures at each stage of delay, time division multiplexers (or demultiplexers) for radar and communication systems can be implemented (7).

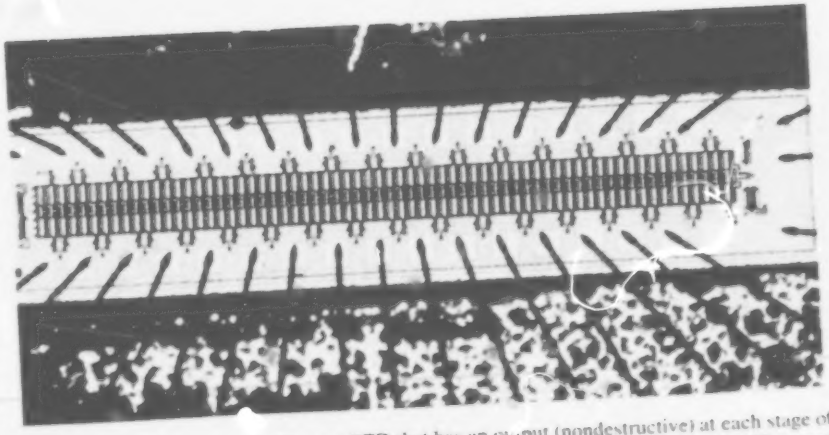


Fig. 6. Photomicrograph of a 32-stage CTD that has an output (nondestructive) at each stage of delay. The total device size including the rectangular bond pads for connecting the 32 output leads is 0.1 by 0.5 cm. The CTD itself is the long rectangular structure between the rows of bond pads. [Courtesy of G. Wecker, Reticon Corporation]

Surface Acoustic-Wave Principles

In SAW devices information is represented by ultrasonic acoustic waves propagating freely along the surface of a planar solid, rather than by charge packets moved along by an external energy source as in the CTD. These waves, whose intensity is greatest at the surface, involve exceedingly minute motions of the particles of the solid on which they propagate: in typical SAW devices the particle displacements are at most a few angstroms (8).

The wave velocity depends on the materials comprising the path of propagation and on the structure employed. The surface wave velocity is independent of frequency for a uniform solid, but propagation on a layered solid is dispersive (frequency-dependent). A typical velocity for surface acoustic waves is 3×10^3 meters per second, which is 100,000 times lower than the velocity of electromagnetic waves in a vacuum; hence, at a given frequency an SAW wavelength (wavelength = velocity/frequency) is about 100,000 times smaller than an electromagnetic wavelength. (For example, a 300-Mhz signal in an SAW device would have a wavelength of only about 10 μm.) The lowest frequency at which SAW devices are useful is limited to about 5 Mhz by the dimensions of available crystals and the increase in wavelength as frequency decreases, while wave attenuation and fabrication difficulties set an upper limit of a few gigahertz for room-temperature devices. Thus the CTD and SAW devices typically operate in complementary frequency ranges, and the data rates of SAW devices may be very high.

It is easy to convert electrical signals to surface acoustic waves (and vice versa) by use of transducers based on the piezoelectric effect. If a surface wave propagates on a piezoelectric crystal (such as crystalline quartz) or on a nonpiezoelectric crystal coated with a piezoelectric layer (such as a silicon wafer coated with piezoelectric zinc oxide), a traveling electric field accompanies the propagating acoustic wave. In effect, the displacements of the atoms of the medium produce local electric fields inside and just outside the solid. One can couple external sources or detectors to these electric fields by means of conducting electrodes evaporated onto the surface and shaped by photolithographic techniques like those used in making integrated circuits.

Figure 3a shows an SAW device on a piezoelectric insulator. The signal source creates electric fields between the evenly spaced fingers of the input electrode transducer on the left side in Fig. 3a, producing surface waves that propagate toward the output electrode transducer (on the right), where they produce an output voltage that might be displayed on an oscilloscope. (Waves launched to the left by the input transducer are absorbed by the absorbing material on the left end of the crystal.) Each finger of the input transducer acts as a source of waves; the frequency of the strongest wave excitation by the input transducer will be that for which the spacing between alternate fingers equals one SAW wavelength, because for this frequency the contributions from the individual electrodes will add in phase. Thus, using the earlier example of 300-Mhz waves and a phase velocity of 3×10^3 m/sec, the center-to-center distance, L, of alternate electrodes should be one wavelength or 10 μm.

If the transducer on the right in Fig. 3a were identical with that on the left, this device would function as a simple delay line, delaying signals by about 3.3 μsec for each centimeter separating the two transducers. The device would operate well only in a band of frequencies near the design center frequency (300 Mhz) because at frequencies far removed from that the waves produced by individual electrode fingers would tend to interfere destructively, causing decreased output. Such frequency-selective transmission has been used to make miniature SAW bandpass filters for operation in television sets and other electronic equipment (Fig. 3b).

Considering again the output transducer as it is actually shown in Fig. 3a, we note that the placement of fingers varies along the transducer, from larger spacings on the left side to smaller ones on the right. This variable spacing causes the left end of that transducer to respond most strongly to lower-frequency waves (larger wavelengths) and the right end to higher-frequency waves. Thus if a brief input signal composed of many frequency components is applied to the input transducer, the low-frequency components produce a voltage at the output transducer sooner than the high-frequency components do because they arrive earlier at the portion of the output transducer that is responsive to them. Thus one can make a frequency analyzer with this simple structure, in which frequency differences in an impulse excitation are converted to differences in arriv-

al time which can be measured on an oscilloscope (Fig. 3c).

Let us consider two other uses of this device, in pulse expansion and pulse compression. If we apply a voltage of very short duration to an input transducer consisting of just one pair of electrode fingers at the left of the device of Fig. 3a, a surface wave having many frequency components is produced, and when this wave passes under the output transducer shown, the output voltage is a waveform with a frequency that varies from low to high (Fig. 3d). Such a waveform is called onomatopoeically a chirp. Chirp waveforms have applications in radar, sonar, medical ultrasonics, and other signal-processing fields. In radar, for example, a reflected signal from a very distant target may be lost in noise because of the limited amount of peak power one can generate and transmit from an antenna; in medical ultrasonics, the instantaneous intensity that can be safely transmitted into the body is limited to avoid tissue damage. One can alleviate these problems by amplifying and transmitting a chirp waveform generated as described above (*9*), receiving the echo (also a chirp that starts at a low frequency and rises to a high frequency), and applying the received echo to the broadband input transducer of an SAW device like that of Fig. 3a but with its output transducer reversed (that is, with its closely spaced electrode fingers nearer to the input transducer than the widely spaced fingers, as shown in Fig. 3e). In this second SAW device the low-frequency components are detected after the high-frequency ones, and the long-duration low-amplitude received signal is compressed into a much shorter pulse of higher amplitude (Fig. 3f). Thus the travel time to a reflecting target could be determined with precision in spite of interfering noise.

These examples show how very simple patterns of electrodes on the surfaces of piezoelectric crystals are used to make SAW devices for such varied operations as signal delay, filtering, frequency analysis, and pulse expansion and compression.

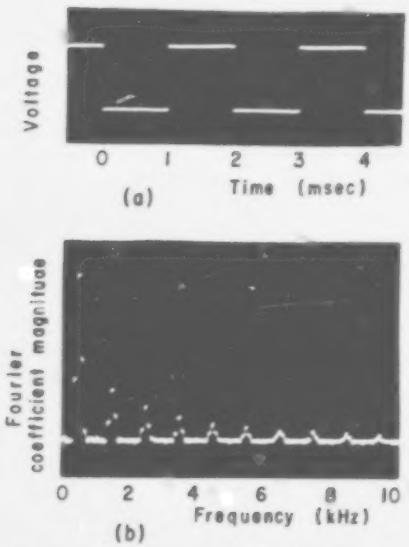


Fig. 7. CTD spectrum analysis: (a) 500-hertz square-wave input, and (b) CTD spectrum analyzer output showing the magnitudes of the lowest-frequency spectral components. The magnitude of the *n*th harmonic is proportional to 1/*n* as expected.

Transversal Filtering

Filters are circuits that are generally used to pass or reject different frequency components, which is probably the most common function performed in electronic systems. It is difficult to conceive of any moderately complex electronic system that does not employ some type of filter (*10*).

One approach to filtering is to use a transversal filter, which is shown in block diagram form in Fig. 4. The output of a transversal filter is a weighted sum of delayed replicas of the input signal. Any desired frequency response can be obtained (*11*) by properly choosing the weighting coefficients W_i (within constraints determined by the allowable amount of delay).

Surface acoustic waves provided the first practical and convenient way to make high-performance, compact transversal filters which can operate at data rates of hundreds of megahertz. Charge-transfer devices have also been used extensively to perform transversal filtering, but at lower frequencies (100 hertz to 5 megahertz). The conventional approach to filtering (inductance-capacitance and active filters) in this frequency range requires the use of circuit components (for example, inductors and precision resistors) that cannot be integrated onto a single IC. This can severely limit the complexity of the system that can be integrated, since a relatively large amount of circuit area is required to interface the IC with the external components. In addition, the use of external components can significantly increase the cost and degrade the reliability compared to those of a fully integrated system.

In the next two sections we give examples of the use of CTD and SAW devices to demonstrate their advantages.

CTD Transversal Filters

There are several methods for implementing transversal filters based on CTD's. The highest performance is obtained with the split electrode approach (*12*), in which some of the metal gates used in transferring the charge are split into two sections, as shown in Fig. 5. The charge that flows into each part of the split gate is proportional to the product of the channel voltage and the gate capacitance. The channel voltage is equal to a delayed value of the input voltage. The capacitance of each side of the gate is determined by the placement of the gaps (Fig. 5), which are located to yield the weighting coefficients W_i. It is therefore only necessary to sense the differential value of the displacement charge between the two sides of the *i*th split gate G_3 in order to obtain multiplication of the delayed input signal by W_i. By summing the charge on the two bus lines that connect the split gates, the sum indicated in Fig. 4 can be obtained.

This approach has the advantage of high density [800-stage transversal filters have been made (*13*)] as well as high performance [10-bit accuracy of the coefficients W_i is possible and filter dynamic ranges in excess of 80 decibels have been achieved (*14*)]. The major disadvantage of this approach is that since the weighting coefficients are fixed once

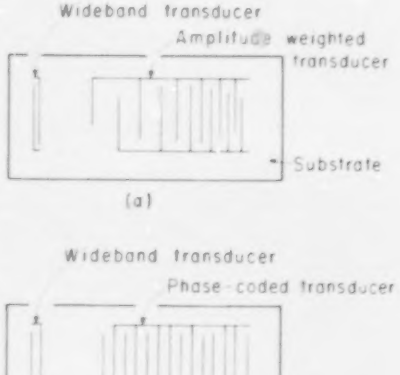


Fig. 8. Tap weighting and phasing of an SAW transversal filter. (a) Schematic illustration of weighting the transducer elements by varying the amount of finger overlap (in this case the middle frequencies are emphasized relative to the low and high frequencies). (b) In the transducer on the right all fingers overlap the same amount but the phasing of pairs of fingers varies because of changes in bus bar connections. A binary sequence represented by transducer phasing is indicated below the output transducer.

the photolithographic mask has been made, it is not possible to vary the filter response after the device is fabricated or to perform adaptive filtering functions. This type of device is limited to applications that require a large number of filters having the same response in order to amortize the cost of mask making over many units.

For lower-volume application a programmable device would be more desirable because it could be tailored by the user. One way to obtain this capability is to sense nondestructively the delayed value of the input at each stage of a delay line, bring this off the IC, and then perform the multiplications and summation externally. A photomicrograph of a commercially available 32-stage tapped analog delay line is shown in Fig. 6. External resistors and amplifiers can be used to perform the multiplications and summation, providing a complete transversal filter whose characteristics (tap weights) can be altered.

It is informative to compare the amount of computation that can be performed by a CTD transversal filter with the capabilities of a general-purpose computer such as an IBM 360 or a CDC 6400. These computers can perform about 10^7 multiplications and sums in 1 second (*15*). A CTD filter of 10^3 stages can perform 10^3 multiplications and 10^3 additions (2×10^3 operations) in parallel during each clock period. Since a filter can be operated as fast as 5 Mhz, this means that a CTD filter can achieve a processing rate of $(2 \times 10^3) \times (5 \times 10^6) = 10^{10}$ operations per second. Thus the CTD filter, which might cost as little as $1 or $2, can perform three orders of magnitude more computation per second than the large, general-purpose computers. Of course these computers are not limited to special-purpose operation like the CTD filters, but because of the parallel architecture of the CTD processor and the efficiency of analog computation, the CTD approach retains enormous processing advantages for special applications.

Spectral or Fourier analysis (the decomposition of a time waveform into its frequency components) could be performed by a bank of CTD band-pass transversal filters. This is very wasteful, however, compared with an approach that makes use of the chirp *z*-transform algorithm (*16*). With this algorithm, an *N*-point Fourier transform could be implemented on a few IC's using only four *N*-stage CTD transversal filters (*17*).

The first 100 coefficients from the output of a 500-point CTD Fourier-spectrum analyzer are shown in Fig. 7b. The input signal was a 500-hertz square wave (Fig. 7a). The harmonics of the square-wave spectrum are clearly evident at 1.5, 2.5, . . . khz. A CTD spectrum analyzer realizes the advantages expected from the use of the parallel analog processing of the CTD transversal filters. It can calculate a spectrum many times faster than the fastest special-purpose digital computer. In addition, because it can be integrated onto a few IC's, substantial savings can also be made in cost, size, weight, and power (*18*).

Spectral analysis by use of the CTD chirp *z*-transform is expected to be very important in image processing, speech recognition, radar Doppler processing, sonar spectral analysis, video bandwidth compression, and other applications requiring low-cost spectral analysis.

Processing Signals with SAW Devices

Surface-wave devices are useful in part because the SAW electrode transducers themselves are transversal filters: delay results from propagation of the wave along the substrate under the electrodes, which tap off a portion of the wave energy, and summing of currents from individual taps occurs in the broad bus bars connecting the fingers. Tap weighting and phasing is easily accomplished by (i) tailoring the amount of overlap of adjacent fingers (the output from each pair of fingers is roughly proportional to the amount of overlap) and (ii) connecting corresponding fingers of each pair to one bus bar or the other for proper phasing (Fig. 8).

Most of the SAW devices (*19*) used to date have been bandpass filters (*20*) based on the transversal filter concept. Illustrative are SAW filters used in connection with television receivers, both in the intermediate-frequency sections of commercial receivers (*21*) and in some electronic television games where they confine the output transmission to the selected channel (channel 3 or 4). Such SAW filters are smaller than conventional filters made with inductors and capacitors, they can be reproduced precisely once the photolithographic master has been designed and made, and they do not require adjustment after manufacture. Another example is the SAW inverse filter (*22*), whose frequency response is ap-

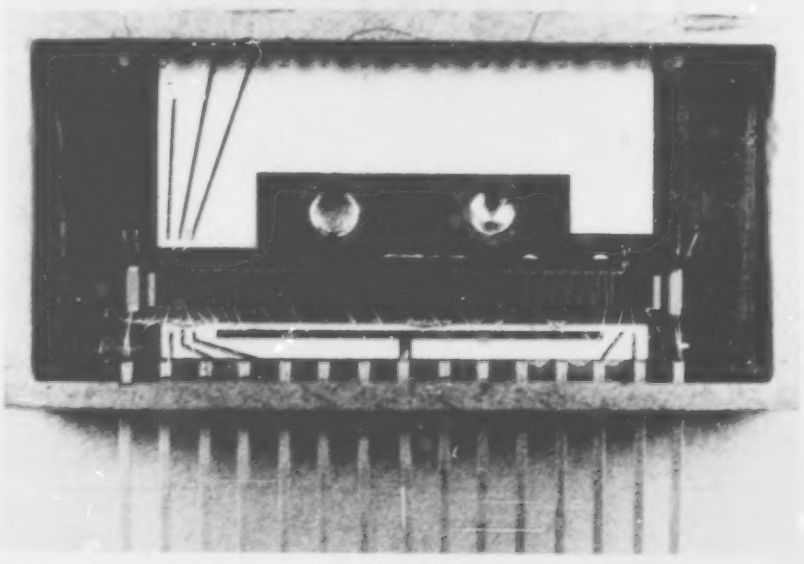


Fig. 9. Electronically programmable SAW matched filter mounted in a conventional flatpack. Thirty-one taps are visible on the ¾-inch-long silicon substrate that extends across the lower portion of the package. Gold wires carry currents from the taps to the summing electrode on the white insulating circuit board below the silicon device. Piezoelectric zinc oxide sputtered onto the silicon after fabrication of the circuit is the light rectangle at the far left side of the wafer; the 100-Mhz SAW input transducer is located there. Tap phasings are set by read-only memory located above the line of taps. The signal to select the desired filter characteristic reaches the device through one of the six angled conductors crossing the white circuit board at the left end of photograph. The device can be used to generate the waveform to which it is matched if the SAW transducer at the right side of the silicon wafer is used. [Courtesy of F. Hickernell, Motorola Corporation]

proximately reciprocal to that of the sensor whose output is filtered. For example, partial correction for the limited frequency response of the source and receiver in a nondestructive tesing or medical ultrasonic system can be achieved with such a real-time filter. The matched filter is another useful device that can be made with the SAW (or CTD) technology. The impulse response, $h(t)$, of a matched filter is the time reverse of the signal, $s(t)$, to which the filter is matched; thus $h(t) = s(-t)$. The matched filter (*23*) has a much larger output for the desired signal $s(t)$ than it does for other input signals because the outputs from all the electrode fingers (taps) add in phase when the desired signal is midway in its passage through the filter. An example of such a matched filter was shown in Fig. 3e; the chirp output from the first SAW device was compressed by the second device because of the reversal of the transducer on the right (*24*). Matched filters can also be used in digital transmission systems for source-receiver synchronization, and in secure communications systems.

Changing the responses of these filters is sometimes of interest; for example, one might want to change the codes to which a receiving filter was matched in a secure communications system. Surface acoustic wave devices that can be programmed electronically in a few microseconds have been realized in several ways by combining piezoelectric and semiconductor materials. One practical approach has been to change the phasings of the taps on SAW filters by providing semiconductor diode switches for each tap to determine which bus bar it is connected to at a particular time (*25*). Field-effect transistors employing the piezoresistive effect have also been used as taps in a matched filter (*26*) constructed in a silicon wafer onto which zinc oxide is sputtered; a programmable read-only memory on the silicon wafer determines the proper phase for each tap when the filter is set to match a particular code (Fig. 9). Another powerful approach to programmability employs nonlinear effects: the input signal and a propagating or stored signal representing the waveform being sought mix in the nonlinear device to produce an output at the sum frequency; the output is maximum when the desired signal is present in the input. Such nonlinear SAW devices can also be used to perform the complex operations of convolution and correlation of high-frequency signals (*27*).

Some idea of the variety of applications to electronics that have been demonstrated can perhaps be obtained from these final examples: SAW's can be used for modulating and deflecting light in integrated optics (*28*); in an undersea pressure sensor where changes in wave travel time between transducers are measured (*29*); for temperature-compensated signal storage and delay in small disk-shaped quartz wafers (*30*); and, through partial reflection of waves at grooves or electrodes, to make high-quality resonators and very selective radio-frequency filters (*31*).

Conclusions

In charge-transfer and SAW devices new electronic principles provide efficient low-cost solutions to existing signal-processing problems. The combination of analog computation with a discrete time signal representation, which is available with CTD signal-processing devices, is found to have considerable advantages in reducing the cost, power, size, and weight of signal-processing systems. An important factor in obtaining these reductions is the compatibility of the CTD technology with standard MOS circuits, which allows integration of additional non-CTD functions and allows the design of integrated circuits that are complete subsystems in themselves. In addition, the ability to vary externally the transfer rate of the charge packet is the basis of many unique CTD applications.

In SAW devices the analog signals are represented by waves moving at a speed that cannot be varied, but the technology permits one to make sophisticated signal processors that operate at very high data rates with signal frequencies up to gigahertz. Taken together, the CTD and SAW technologies permit one to realize signal-processing functions that a few years ago would have seemed impossible.

References and Notes

1. An excellent general review of CTD's and of the literature before 1975 is given in C. H. Sequin and M. F. Tompsett, *Charge Transfer Devices* (Academic Press, New York, 1975). More recent articles are available in the special issue on charge transfer devices, *IEEE J. Solid-State Circuits* **SC-11** (February 1976).
2. There are two basic types of CTD: the charge-coupled device and the bucket brigade device. They are so similar with respect to the applications discussed in this article that they can be considered simultaneously.
3. T. E. Linnenbrink, M. J. Monahan, J. L. Rea, in *Proceedings of the International Conference on the Applications of Charge-Coupled Devices*, J. Lignado, Ed. (Naval Electronics Laboratory Center, San Diego, 1975), p. 443.
4. W. J. Hannan, J. F. Schanne, D. J. Waywood, *IEEE Trans. Mil. Electron.* **MIL-9**, 246 (July 1965).
5. L. J. M. Esser, *IEEE ISSSC Dig. Tech. Pap.* **18**, 28 (February 1974).
6. M. F. Tompsett and E. J. Zimany, Jr., *IEEE J. Solid-State Circuits* **SC-8**, 151 (April 1973).
7. T. F. Cheek, Jr., A. F. Tasch, Jr., J. B. Barton, S. P. Emmons, J. E. Schroeder, in *Proceedings of the International Conference on the Applications of Charge-Coupled Devices*, J. Lignado, Ed. (Naval Electronics Laboratory Center, San Diego, 1973), p. 127.
8. The properties of surface acoustic (or elastic) waves have been discussed in many publications; for example, see I. A. Viktorov, *Rayleigh and Lamb Waves: Physical Theory and Applications* (Plenum, New York, 1967); G. W. Farnell, *Phys. Acoust.* **6**, 109 (1969); R. M. White, *Proc. IEEE* **58**, 1238 (1970).
9. For a comprehensive survey of SAW devices employing the chirp principle, see J. D. Maines and E. S. Paige, *ibid.* **64**, 639 (1976).
10. D. E. Johnson, *Introduction to Filter Theory* (Prentice-Hall, Englewood Cliffs, N.J., 1976).
11. L. R. Rabiner and B. Gold, *Theory and Applications of Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, N.J., 1975).
12. D. D. Buss, D. R. Collins, W. H. Bailey, C. R. Reeves, *IEEE J. Solid-State Circuits* **SC-8**, 138 (April 1973).
13. C. R. Hewes, in *Proceedings of the International Conference on the Applications of Charge-Coupled Devices*, J. Lignado, Ed. (Naval Electronics Laboratory Center, San Diego, 1975), p. 170.
14. R. W. Brodersen, C. R. Hewes, D. D. Buss, *IEEE J. Solid-State Circuits* **SC-11**, 75 (February 1976).
15. R. Turn, *Computers in the 1980's* (Columbia University Press, New York, 1974), p. 75.
16. L. R. Rabiner, R. W. Shafer, C. M. Rader, *IEEE Trans. Audio Electroacoust.* **AU-7**, 86 (1969).
17. R. W. Means, D. D. Buss, H. J. Whitehouse, in *Proceedings of the International Conference on the Applications of Charge-Coupled Devices*, J. Lignado, Ed. (Naval Electronics Laboratory Center, San Diego, 1973), p. 95.
18. D. D. Buss *et al.*, in *ibid.* (1975), p. 267.
19. Surface acoustic wave devices and applications have been reported primarily in the *Ultrasonics Symposium Proceedings* (yearly) and the *IEEE Transactions on Sonics and Ultrasonics*. Interested readers should also consult the special issue on surface acoustic waves, *IEEE Proc.* **64** (May 1976), and J. de Klerk, *Phys. Acoust.* **11**, 213 (1975).
20. For example, see R. M. Hays and C. S. Hartmann, *Proc. IEEE* **64**, 652 (1976).
21. A. J. DeVries and R. Adler, *ibid.*, p. 671.
22. G. L. Kerber, R. M. White, R. W. Wright, in *1976 Ultrasonics Symposium Proceedings*, J. de Klerk and B. McAvoy, Eds. (Institute of Electrical and Electronics Engineers, New York, 1976), p. 577.
23. G. L. Turin, *IRE Trans. Inf. Theory* **IT-6**, 311 (June 1960).
24. Pulse compressors for radar applications have been made (by somewhat different SAW techniques) with low spurious signal levels and time-bandwidth products as high as 10^6; see R. C. Williamson, *Proc. IEEE* **64**, 702 (1976).
25. J. de Klerk, *Phys. Acoust.* **11**, 213 (1975).
26. F. Hickernell, M. Adams, A. London, H. Bush, in *1975 Ultrasonics Symposium Proceedings*, J. de Klerk and B. McAvoy, Eds. (Institute of Electrical and Electronics Engineers, New York, 1975), p. 223.
27. G. S. Kino, *Proc. IEEE* **64**, 724 (1976).
28. E. G. H. Lean, J. M. White, C. D. W. Wilkinson, *ibid.*, p. 779.
29. T. M. Reeder and D. E. Cullen, *ibid.*, p. 754.
30. I. M. Mason, E. Tsizdofrangakis, J. Chambers, *ibid.*, p. 610.
31. D. T. Bell, Jr. and R. C. M. Li, *ibid.*, p. 711.

New Memory Technologies

Jan A. Rajchman

The multifarious capabilities that have made the computer the great success of our age are due to exploitation of the high speed of electronic computation by means of stored programs. This process requires that intermediary results be stored rapidly and furnished on demand for long computations, for which high speed is worthwhile in the first place. Storage devices or memories must have capacities sufficient not only for the intermediary results but also for the input and output data and the programs. The demand for fast access and large capacity has grown constantly. Today there are memories accessible in tens of nanoseconds and memories with more than a billion bits, but in general fast memories have small capacities and memories with large capacities have slow access.

Ideal would be a single device in which vast amounts of information could be stored in nonvolatile form suitable for archival record-keeping and yet be accessible at electronic speeds when called for—perhaps a shoe-box device containing 10^{12} bits accessible at random in nanoseconds. So far there is no way to realize this ideal. Fortunately, the benefits of large capacity and rapid access can be obtained by use of a hierarchy of different types of storage devices of decreasing capacity and increasing speed. The main hierarchy today comprises, on one hand, large-capacity magnetic recording devices, which are accessed mechanically and serially (reels of tapes, disks, and drums), and on the other hand, fast electronic memories (the core memory and various types of transistor memories). The user may be unaware of the hierarchy, as the distribution of information between the various devices is built and programmed into the system through remarkable software artifices in computer architecture, memory organization, and internal operating programs. An outstanding shortcoming of the present state of the art is the gap (*1*) between the mass-memory storage devices and the electronically addressed memories. The spanning of the gap by software is extremely clever—but it entails not only labor and complexity but elaborate internal programs that preempt much of the memory capacity and tend to slow the processing.

This article deals with electronic memories and new developments in this area (*2*, *3*). In particular, the challenge of developing an electronically addressed memory that fills the gap between the present mechanically accessed magnetic disk memories and fast transistor memories is considered.

Core Memory

One of the first electronic memories was a circulating delay line, a signal transmission device in which the output, properly amplified and shaped, was fed back into the input. Although it was economical, it had the inherent drawback of serial access: the greater the capacity, the longer the average access time. What was really needed was selective access to any stored data in a time that was both as short as possible and independent of the data address or any previous access. This is known as random access, so named to emphasize the total freedom of accessing and therefore of branching (following one or another part of a program), which is indispensable in the execution of stored programs. The first random-access memories (RAM's) were electrostatic storage tubes. In the early 1950's the core memory (*4*, *5*) replaced these early devices, providing a solution to the need for random access that truly fired the emerging computer industry into its fabulous growth.

A core is a tiny ring-shaped piece of magnetic material in which a bit of information is stored by magnetizing it in a particular direction, that is, to one or the other of its two remanent states. The core memory consists of an array of cores strung on fine wires. To retrieve information from a core a current is sent through it that causes it to be magnetized in a given direction that is the same as or opposite to the direction in which it had been left. If the magnetic state changes, a voltage is induced that provides the readout and at the same time a signal for restoring the core to its original state. To select one core in an array of many, two current pulses are applied simultaneously to one row and one column of the array, and the core responds only to the coincidence of the two excitations. This is possible because the ferrite cores used have a square hysteresis loop and are very uniform. Individual cores are made and sorted automatically, hundreds per second, and then are strung on wires, generally manually. During about a decade (1955 to 1965) many interesting magnetic structures were developed in which the elements were made by a batch process rather than handled individually (*6*). It turned out, however, that the ratio of electronics to magnetics in the system had a far greater influence on performance and cost than the nature of the magnetic structure. Hence, the established and constantly improving core technology prevailed.

The core memory has become the main internal computer memory and was used universally until challenged recently by semiconductor memories. Typical are memories with 1 million words of 30 to 60 bits each, randomly accessible in 1 microsecond. The core memory has also been extended to very large capacities, of the order of 100 million words, and has proved the benefits of filling the gap in large systems, that is, software simplification and more efficient computation (*6a*). However, this gap-filling solution is too expensive (about 1 cent per bit) to be of general use. The core memory is still dominant today. From 1966 to 1971 the annual production of cores in the United States increased from 10 to 100 billion, and it has probably again increased tenfold from 1971 to 1976. The market outside IBM is estimated to be more than 300 billion cores.

Semiconductor Memories

Since the mid-1960's the spotlight has shifted from the core memory to semiconductor memories. This has resulted from the development of large-scale integration (LSI) techniques for silicon that permit the mass fabrication of microscopically scaled arrays of transistors and all their interconnections through a single set of batch processes. This revolutionary development has had a profound influence on the whole elec-

The author is retired vice president for information sciences, RCA Laboratories, Princeton, New Jersey 08540, and visiting professor at the Electrical Engineering and Computer Science Department, University of California, Berkeley 94704.

tronics industry and is the subject of another article in this issue (7).

It took many years for the semiconductor industry to reach its present proficiency. The first impact on memories was the replacement of tubes by transistors in circuits associated with core memories, which resulted in savings in power, space, and eventually cost. This led to modern core memories with high ratios of electronics to magnetics—and incidentally to general knowledge of memory systems that is still basic today. The first integrated semiconductor memories appeared in the mid-1960's (8). Early transistor memories used static cells with bistable states, each made of a flip-flop circuit with two branches so connected that the "on" state of one ensured the "off" state of the other. At first, relatively high power (about 1 milliwatt) bipolar transistors were used, with which very short access time (less than 100 nanoseconds) could be reached. Hence bipolar memories were used in conjunction with cores, as fast "scratch pads," a hierarchy concept that has been expanded in today's architecture, where fast "cache" memories are used with slower main semiconductor memories. However, the real impetus in the area of semiconductor memories came with the development of unipolar field-effect transistors of the metal oxide semiconductor (MOS) type, which made possible low-power cells (typically 50 microwatts today). In bipolar transistors conduction is by both electrons and holes, whereas in MOS's it is either by electrons (*n*-type transistors) or by holes (*p*-type transistors). (Complementary MOS's, however, are made of both *p*-type and *n*-type transistors.) The LSI technologies for the two involve similar processing and pattern-forming steps. In general, Si-polars are used for high-speed and MOS's for high-capacity memories.

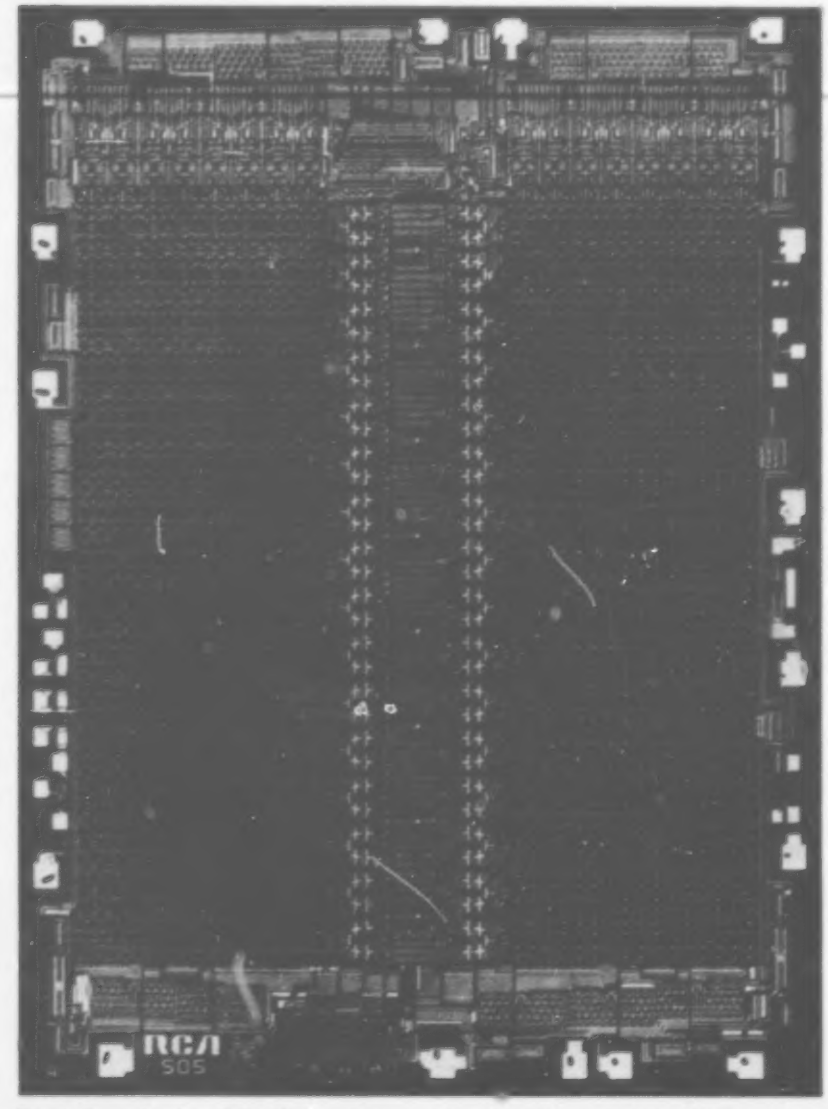


Fig. 1. Example of a 1K RAM on a chip measuring 37 by 51 mm. This RAM has static six-transistor cells, an access time of about 100 nsec, and a total power of 35 mw. High speed and low power in its MOS circuits are achieved through the silicon-on-sapphire (SOS) technology, in which silicon is grown on a sapphire substrate in transistor areas only so that the connecting leads directly on the sapphire have a very low capacitive loading. [Source: RCA Corporation]

In LSI technology the active transistors and all their connections are simultaneously made by unified batch processes. Consequently, the larger the chips cut from the original silicon wafer (which is 5 to 10 centimeters in diameter) the greater the economy, since there are relatively fewer costly connections to the outside. On the other hand, the difficulty of making perfect chips grows with their area. An optimum size turns out to be about 7 to 8 millimeters on the side, for which the yield is only about 20 percent. This puts a great premium on the effective use of the silicon area and has led to a complete reversal of emphasis in circuit design. Whereas efforts were made previously to minimize the active devices (tubes and later single transistors), in LSI the transistors occupy a small area compared to that needed for connections and their spacings, hence ingenuity in design is centered on minimizing requirements for connections and simplifying geometrical layouts (see Fig. 1).

For memory cells, one figure of merit is the bit area, kW^2, where k is a measure of cell complexity and W is the width of line or space attained with a particular technology (W is typically 5 micrometers). Another is the number of lines and contacts necessary per cell (9). Static bipolar cells show low merit under these criteria, as they require two or more transistors and two resistors. The same is generally true of static MOS cells made of *n*-type or *p*-type transistors or of both types in complementary symmetry. These MOS cells have various relative advantages, but all require four to ten transistors. It was found that much better merit factors are achievable in dynamic cells, in which the bit is stored on a capacitance and is periodically refreshed. At first, three MOS transistors were needed for these cells, but modern designs do with one. Dynamic cells occupy an area of only about 25 by 25 μm. This dense bit packing is obtained at the expense of a relatively small lengthening of average access time, which is taken up by the refresh cycle. However, because of the need to refresh information, the dynamic cell favors periodically timed or

synchronous systems, which are less flexible than asynchronous systems. Hence the static cell with its uncompromising performance has not lost out.

In general, a memory chip carrying N bits is organized $N \times 1$, that is, it has one read-write bit channel and N addresses. Usually N is a power of 2, such as 1024, 4096, or 16384, in which cases the chip is referred to as 1K, 4K, or 16K. Internally, the N storing cells are physically arranged in an array and are selected by the two array lines on which they lie. The chip carries two decoders that select these two lines from a binary code. Hence the number of address lines to the chip, $\log_2 N$, is relatively small (10 for 1K, 12 for 4K, and 14 for 16K). The chip also carries read-write circuits, timing circuits, and regeneration circuits when dynamic cells are used. These peripheral circuits preempt two-thirds of the chip area in a typical MOS RAM 4K chip. Chips may be packaged in separate cans, which typically have 16 pins, and the cans may be mounted and interconnected on printed boards. Alternatively, chips can be mounted directly on the board.

Whatever technique is used for mounting chips, the interconnections are the dominant cause of failure. The mean time between failures (MTBF) per component is typically more than 1000 years, after an initial weed-out period of 10 to 200 hours during which it is not unusual for 1 percent of the components to fail (9). Despite this long component MTBF, a system with hundreds of thousands of components could well have an MTBF below several weeks. Fortunately, error-correcting codes with redundant additional digits can generally increase the MTBF; in fact, the probability of error can be reduced almost at will with a sufficient increase in system complexity. Single-bit errors, which are the major type by deliberate system design, can usually be corrected by relatively simple codes that entail a modest system cost increment (less than 15 percent).

To complete the memory system, of course, power supplies, containing boxes, and coupling and checking circuits are necessary. The resulting system per bit cost is in general almost ten times greater than the chip cost per bit.

The main memory of the IBM 370/168 may be taken as illustrative of the state of the art: it has a capacity of 64 million bits (64 Mb), uses 32,000 2K chips with static cells, has a cycle time of 2 μsec, costs 2.4 cents per bit, and achieves an MTBF of 10^5 hours (with error-correcting codes). Also standard in the last few years have been 4K, 16-pin dynamic cell packages with access times of about 200 to 300 nsec for 0.2 to 0.4 cent per bit (10). This year 16K RAM's have been appearing, and—which is indicative of the rapid progress being made—they have the same speed and power dissipation ($\frac{1}{2}$ watt) and pin count as the 4K RAM's (11). Another indicator of progress is a recently announced (12) bipolar technology known as integrated injection logic (I^2L), which has been used to make one-transistor dynamic cells on 4K chips that outperform their MOS counterparts.

Semiconductor RAM's may well fill the gap between present-capacity RAM's and mechanical disks by simply extending their capacity. A good case for this is made by Hodges (9), who envisages that in 5 years quarter-billion-bit memories will be available that are accessible in 2 μsec and cost 0.04 cent per bit. He assumes a reasonable extrapolated improvement in yield but no deviation from conventional technology. Other extrapolations confirm the increase in capacity and project even greater cost reductions.

Fig. 2. A 16-kb block-addressed charge-coupled memory device. Each of the four 4K chips is organized in a series-parallel-series array and carries its own decoders. Operated at a data rate of 1 Mhz, the mean access time is 2 msec and the on-chip power dissipation only 1.5 μW per bit. The memory is fabricated by a simple MOS n-channel technique (17). [Source: Bell Telephone Laboratories]

At the heart of LSI techniques is photolithography, which makes it possible to conveniently create all the necessary microgeometrical patterns, including the delineation of areas for doping and for metallization. An arsenal of photographic reducing techniques, photoresistive coatings, and etching techniques has evolved. Diffraction of light at visible wavelengths limits the width of well-defined lines to about 2 μm, although at present widths of 4 to 5 μm are typical. A dramatic reduction to submicrometer dimensions can be achieved through the use of electron beams rather than photolithography. An electron beam focused to a fraction of a micrometer traces the desired pattern under computer control, either directly on the wafer or on a mask. The mask pattern is then replicated by x-ray exposure. A great deal of work and capital have been devoted to electron-beam lithography (13, 14), which already provides greater reproducibility in masks of conventional scale and in future will undoubtedly be producing scaled-down structures with high bit densities. Experimental structures with line widths of less than 1 μm have been demonstrated for some time.

Charge-Coupled Devices (CCD's)

One way to increase storage density is to avoid having a direct contact to each storage location, by shifting packets of charge at the silicon oxide interface in MOS devices, as was first described by Boyle and Smith in 1970 (15). Charge, injected or not at the beginning of the line, is shifted by applying phased pulses to the clocked electrodes so that the pattern of presence and absence of charges appears at the end of the line. Here it is amplified and reinjected at the beginning, making out of the CCD shift register line a serial memory with N circulating bits, much like the early delay lines. This return from RAM's to the older concept with its long latency time is motivated by the higher bit densities obtainable. As it turns out, the rules of design of RAM and CCD cells are the same, and there is not much to be gained with CCD's, whose register lines have to be spaced in the same way as the cells of RAM's and whose elements occupy roughly the same length along the line. However, with CCD's it is possible to use a much smaller charge with smaller elements in a scaled-down structure, be-

cause the stored charge is available at the end of the line to produce the readout signal, whereas in RAM cells it is divided by the capacitive loading of other cells on the line. This also leads to simpler readout circuits, and as the access circuits are inherently simpler the peripheral circuits preempt only half rather than two-thirds of the chip area. The net result is a fourfold increase in bit density, a 16K CCD being made on the same size chip as a 4K RAM.

To mitigate the latency time problem, a number of loops are organized in various ways. Each loop can carry its own regeneration, and the loops can be randomly addressed. Or the loops can share their regeneration circuits and be arranged in a series-parallel-series combination with short and long loops. Shared circuits with random loop selection are used in line-addressable RAM's (Larams), taking advantage of some inherent storage in unclocked CCD's. With these arrangements any block of information can be addressed randomly, but the information within the block flows serially, albeit at very high bit rates. While various trade-off's of latency time and complexity can be obtained, the average access time is several hundred microseconds.

There has been intensive work on CCD's (*15–18*) since their inception, and in 1975 and 1976 four manufacturers announced 9K, 16K, and 65K memories (see Fig. 2). Several more 16K CCD's are due in 1977. The CCD's benefit from all the advances that have been made in MOS technology, and hence their great bit density immediately translates into a correspondingly lower cost. On the other hand, their access times are almost 100 times longer than those of RAM's, and they are constrained by their particular data block size and organization and hence do not enjoy the great versatility of RAM's. It is likely that both CCD's and RAM's will be used as gap-filler technologies in the near future.

Whatever their success as memories, CCD's have many other applications well suited to their serial operation. An important area is signal processing. With elements sensitive to light, CCD's become a solid-state TV camera that already has characteristics rivaling those of its vacuum tube vidicon counterpart.

Bubble Memories

The control of cylindrical domains or bubbles in certain magnetic materials, first described in 1967 by Bobeck (*19*), has led to extremely interesting possibilities for memories. The bubbles are formed in a platelet by applying a biasing magnetic field normal to its surface. For a certain range of bias, cylindrical domains of reverse magnetization are formed. The domains can be moved in the plane of the platelet by relatively small field gradients. To produce a shift register, these gradients can be created by the combination of a rotating magnetic field over the whole platelet and an overlay of suitably patterned permalloy. The pattern is usually a succession of T-bars and I-bars, which become tiny magnets that periodically attract and repel bubbles and cause them to propagate along the line of bars. By connecting the two ends of a shift register a storage loop is obtained. At some location in that loop there is a tiny looped conductor, which when energized in one polarity generates a bubble and when energized in the opposite polarity annihilates a passing bubble if there is one. In this way a pattern of bubbles or no bubbles can be made to circulate and be stored in the loop. To read out the stored pattern, at some location in the loop there is a bubble replicator that creates two bubbles out of a passing one, one which continues in the loop and another one which is steered into the detector. Here it is expanded and its presence is sensed by the magnetoresistance of the permalloy guiding structure itself (see Fig. 3) (*20*).

Fig. 3. Photograph of a 66-kb magnetic bubble chip (4 by 5 mm) organized in major and minor loops and having 3-μm bubbles. The large black rectangular area is a chevron expander-detector; here the bubbles are greatly expanded and cause an appreciable change in electrical resistance of the permalloy magnetic guiding structure. (Source: Bell Telephone Laboratories)

Fabrication of a bubble device starts with a substrate of gadolinium gallium garnet (G^3). On these G^3 wafers, which are 5 to 7.5 cm in diameter, an epitaxial layer of doped yttrium-iron garnet (or more recently garnets with calcium and germanium) is grown. Next, an overlay of permalloy and conductors, patterned by standard photolithography, is applied on the wafer. Finally the wafer is cut into chips. Bubble devices require similar but fewer and simpler manufacturing steps than integrated semiconductor devices. A barium ferrite plate provides the necessary biasing permanent field and preserves on the system level the inherent nonvolatility of the bubble memory. Bubble chips can be packed into standard packages, which also hold the necessary magnetic-field generating coils and bias magnets as well as shields from unwanted external fields (see Fig. 4). A 32-pin package has been designed that houses four chips with a total capacity of 280 kilobits (kb).

Because bubble devices are inherently serial, various memory organizations with many storing loops are used, which resemble the CCD organizations that were developed later. Long registers, major and minor loop arrangements, and decoders correspond to CCD serpentine shift registers, series-parallel-series registers, and Larams. Much of the address logic switching necessary within these composite organizations is obtained with the bubble technology itself. Advantage is taken of the local fields created by either a current loop or a bubble that influences the path another bubble will take. Therefore, while semiconductor circuits are still necessary to amplify the sense signals and furnish the inputs and the currents for the rotating magnetic field, they represent a significant but not a controlling economic factor.

A decade of intensive research and development work on bubbles in many laboratories throughout the world has resulted in a basic knowledge of the physics and materials, as well as many ingenious devices, in what has been one of the most important innovations in magnetism (*20, 21*). Large-capacity memories are in pilot production, and it is likely that several manufacturers will announce products in 1977. Chips are likely to carry 32 to 100 kb. Bubble memories intended mostly for gap-filler applications are called files, and they clearly compete with CCD's. Their chief disadvantage is a bit rate of only several hundred kilobits per second, compared to several megabits per second in CCD's. Their chief advantage is nonvolatility and added reliability, although the latter point is being debated. While they might be less expensive because of their simpler processing, this is an uncertain advantage in view of the experience

and volume production achieved in the manufacture of CCD's, which have a head start of several years.

Recent laboratory developments may give bubbles a greater competitive edge. With the minimum photolithographic line width of 2 μm, the permalloy bar file (PBF) used so far has a period of 28 μm and 5-μm bubbles. This has been bettered by a newly demonstrated pattern of asymmetrically shaped disks with a period of 16 μm, 3-μm bubbles, and a chip density of 68 kb (*20*). Extension to 8-μm periods and 250-kb chips is foreseen. Another ingenious innovation, the contiguous disk file, has a gapless structure and tolerates a relatively coarse photolithography; in this file 1-μm bubbles have been propagated along the periphery of contiguous 5-μm disks (*21*).

Perhaps the most interesting development is the bubble lattice file, in which bubbles are packed in a hexagonal lattice determined by bubble-bubble interactions and the densities achieved are four times greater than those of a PBF employing bubbles of the same size (*22*). Information is no longer coded by the presence or absence of a bubble. Instead, use is made of the fact that the magnetization in going radially out of the bubble reverses by rotation within the domain (Bloch) wall, and that the sense of that rotation can be either the same all the way around the bubble or else in one direction in one half and the other in the other half. Entire lines of the lattice can be translated by electrical conductors parallel to the lines but spaced many lines apart. An extracted line at the end is read out seriatim, as in PBF's, by taking advantage of the fact that bubbles in one state move along a field gradient and those in the other state move normally to it.

While their low bit rates and debatable economy make bubble files of current design marginal competitors as gap-filling memories, there is great promise in the newest developments. Bubbles have other applications. For instance, their serial access and their ability to store without holding power make them a natural device for recording audio signals, since storage capacities for minutes of recording seem already economically feasible. The same properties render them valuable in space applications. There are also various uses in telephone systems.

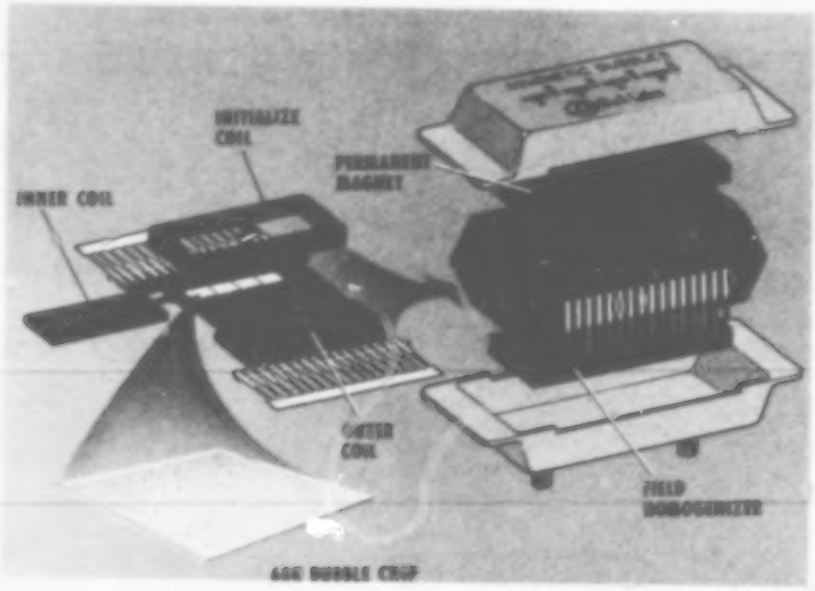


Fig. 4. Scale drawing of a magnetic bubble package. The four-chip package stores 270 kb and contains the necessary biasing permanent magnet and coils for the rotating field. (Source: Bell Telephone Laboratories)

Electron Beam–Accessed Memories

The electron beam is an addressing pointer of high definition and energy density that can easily be deflected. In storage tubes of the 1940's there were severe limitations to such addressing because of the use of surface charge storage and inadequacies in focusing and deflecting the beam. Two recent innovations, storage within a semiconductor and compounded deflection, may bring us closer to realizing the inherent potential of beam addressing. Two groups, one at General Electric working on the beam-addressable MOS (Beamos) (*25*), and the other at Micro-Bit Corp., have memory systems nearing production. For example, consider the Beamos tube with 32 Mb.

The target of this tube is a silicon wafer that has no geometrical pattern. On a p-type substrate there is an epitaxially grown n-type layer, then a thin layer of silicon dioxide, and finally a thin layer of aluminum. The thicknesses are such that a 10-kev electron penetrates through the metal and SiO_2 layers and generates electron-hole pairs in the n-layer. In order to write, a positive bias is applied to the aluminum layer, and holes that are generated in the SiO_2 are trapped near the Si-SiO_2 interface. To read out, the element is interrogated with a beam of lower intensity and no bias voltage. Many electron-hole pairs are produced in the n-layer, and if there are trapped charges many holes diffuse to the n-p junction before recombining. Here they are detected as a current in the external circuit. If there are no trapped charges, the holes recombine and produce no output. The output is typically 1000 times greater than the beam current because each high-energy electron generates many electron-hole pairs. Some erasure occurs on readout, and after a number of cycles the information must be refreshed. (However, with the beam off, charge will remain for as long as 1 month.) Thus the silicon target provides storage in depth, not on the surface, as well as local amplification of the signal. Its only drawback is a fatigue effect that reduces the readout, and this may be cured by relatively infrequent annealings of the tube.

The addressing is in two parts. First, the beam is deflected by a short conical structure of low aberration and strikes normally one of the apertures of a matrix of lenslets. The matrix is made up of two metal plates that have an array of holes (an 18 by 18 array on 1.5-mm centers) and are maintained at different potentials. Each precisely aligned pair of holes forms an immersion lenslet that focuses the beam to 2 to 4 μm. Second, the beam is deflected by bars running along rows and columns between the holes of the matrix. No matter which lenslet is reached, the reduced beam will be subject to the second deflection. In this compounded deflection the accuracy and stability at each step need only be a small fraction of what would be required with a single step. Data can be laid out at will within a lenslet field, and data blocks can vary over wide limits. However, guard bands are required between blocks and lenslets fields. In practice, the best operation is achieved with random access to a block of information and serial access

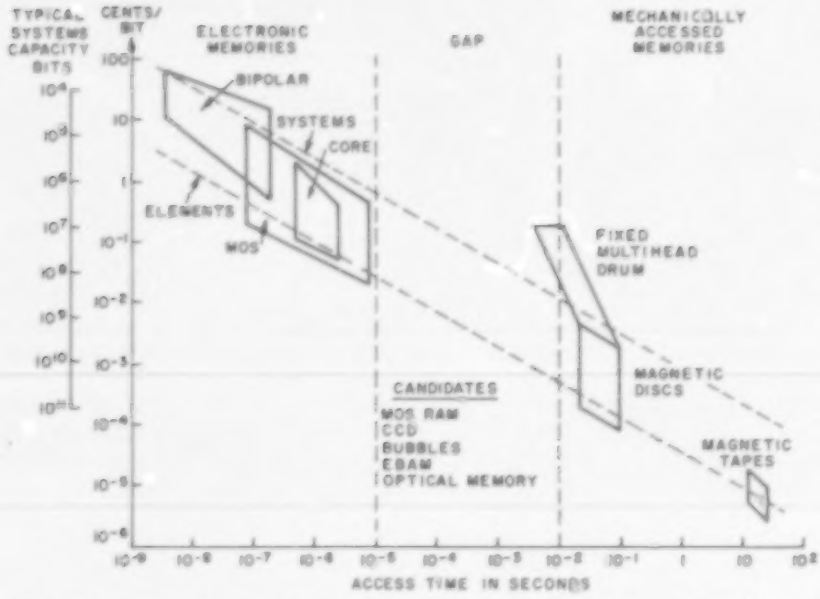


Fig. 5. Cost per bit and capacity plotted against access time for memories in use today. Costs at the element and system levels are indicated by dashed lines.

within the block. The access time is typically 30 μsec and the bit rate 30 Mb/sec.

In a system of large capacity a number of tubes can be configured in various ways. Tubes can be addressed singly and many circuits shared between them for low cost. Alternatively, many tubes can be addressed in parallel for high data rates.

The resort to vacuum technology has clear benefits: it eliminates all geometrical steps from silicon processing, and it leads to very high bit densities. A bit spacing of 4 μm has been demonstrated and it may be possible to bring this down to 1 μm. Costs should be lower than with LSI technology, which is encumbered with complex patterns and a multitude of connections. On the other hand, the tubes are complex, prone to fatigue, and require a high-voltage supply, and the analog circuits needed for deflection are not as foolproof as digital circuits. Despite many years of work and extraordinary achievements, this technology has not appeared on the market and its future is still uncertain.

Optical Memories

Modern developments in optics such as the laser, holography, and various electrooptical effects, have led to new technologies for information storage that are having success in mass storage devices involving mechanical motion (*26*). Of interest here is the unique possibility they offer for making a single memory with the capacity of disk memories and the accessibility of transistor memories (*27*). Briefly, the idea is as follows.

A laser beam is deflected in two directions and passes through an array of light valves in such a way that the image of the array is focused on a storage medium at some location *x,y*. Some of the light from the laser bypasses the array and strikes the medium at the same location *x,y*. Thus a hologram is produced that stores the pattern of openings of the light valves, or one page of information. Because of the redundancy inherent in holograms, the medium does not need to be perfect at the microscopic level of every bit, but merely at some reasonable fraction of a page area. To read from a stored hologram, a reference beam is again directed at it. The real image that created the hologram reappears on the array of light valves. In the array, in addition to a light valve at each point, there is also a light sensor and a semiconductor flip-flop or one bit of a transistor memory. The array, or "latrix," can be considered as the main inner transistor memory of the computer (or part thereof), and the whole arrangement as a way of "photographing" the contents of the internal memory on the mass storage medium and vice versa. In this photographing, light transmission is exploited to transfer information over a very wide channel. Many thousands of bits travel between the storage medium and the transistor memory—all at once. This is very important, as this is the first concept for mass storage with direct coupling to the inner transistor memory which thereby avoids the fantastic speeds necessary to bridge the gap to conventional mechanically accessed memories over narrow channels, and provides a whole system with true random-accessibility.

A feasibility model (*28*) has shown the validity of all the principles of the concept: access by compounded selection of page composition and location, wide-channel optical information transfer, and real-time writing and reading of holograms. In addition, great progress has been made in many laboratories in light deflection (by sonic waves and electro-optic effects) and page composers (based on electro- or magneto-optic effects or the distortions of a thin metallic membrane), as well as the appropriate optical and electronic systems. This progress has shown that fairly reasonable holographic systems could be designed. The main difficulty is the storage medium.

There simply is not a storage medium that is sufficiently sensitive to work at reasonable speeds with reasonable laser powers. Many materials were explored. Curie-point writing in thin films of materials such as manganese bismuth provided fine holograms but required inordinate power. More promising is lithium niobate, which has a unique combination of semiconductor and electrooptical effects. Its sensitivity was increased 10,000-fold in recent years, but it needs to be further increased by a factor of 100 to 1000 to be practical. The difficulty can be appreciated by considering that the sensitivity needed is comparable to that of conventional silver halide photography, which depends on a chemical gain mechanism for its high sensitivity. This suggests the possibility of a storage medium with local amplification, where light would merely trigger an effect energized by an auxiliary power source. In fact, devices made of a sandwich of a photoconductor and an elastomer, working in this manner, exhibited good sensitivities. However, these devices suffer from great technological difficulties.

Concluding Remarks

Faster stores cost more and consequently have smaller capacities. A convenient, although oversimplified, way to illustrate the hierarchy is to plot cost per bit and capacity in typical systems in use against average access time, a mixed parameter that takes into account random-access time and bit rates. The plot (Fig. 5) shows a speed gap of three orders of magnitude between electronically and mechanically addressed memories, and the overlapping areas of usage of

bipolar, MOS, and core memories. It also shows that there is usually an order of magnitude difference between costs at the element and the system levels.

How is the gap likely to be filled? The case for LSI silicon technology is very strong, and the revolutionary nature of that technology cannot be overemphasized. In the first place, there are fundamental reasons why it is ideal for memory matrix arrays. In any such array the peripheral access and signal-amplifying circuits will inevitably be made out of transistors, so that also using transistors for the array avoids the complications of heterogeneous technologies and the many connections needed between them. More significantly, such an array is nearly ideal in the sense that it is connection-dominated, the active storing cell occupying a negligible fraction of the area and contributing negligibly to failures. It is therefore difficult to imagine any other array randomly addressed through connections that can compete with LSI RAM's. The LSI technology also offers another good contender, the CCD, that achieves denser and hence cheaper elements, although with some sacrifice in accessibility.

However, the effects of LSI technology lie beyond these considerations. Throughout the history of computers it was considerably more expensive to manipulate bits for computing or logic switching than for storage. A core cost a few cents, but a tube and its components a few dollars. Central processing units with large memories were a natural architectural consequence. Today, for the first time, logic and storing costs are comparable. Thus it makes sense to associate much more logic with memory, which is precisely what is done in microprocessors. Microprocessors may change our thinking about computer system organization and computing methods. For example, the practice of multiprogramming with a large computer may be replaced by single programming by many small computers. In these smaller computers electronic memories may have sufficient capacity, and there may be much less emphasis on the gap. Microprocessors have also given a new dimension to the intriguing possibilities of simultaneous programming by many computers.

To sum up, the LSI technology is the prime contender for filling the gap. It has also brought about microcomputers, which may change our views on the storage hierarchy itself. The bubble technology can provide sufficient capacity, but its economic advantage has yet to be proved on the market and it has performance limitations with respect to LSI RAM's and CCD's. However, it has a great potential for improvement. Despite great progress and further possibilities, electron beam–addressed memories have a difficult road to the market. Optical techniques offer the best solution conceptually, but much research is necessary, particularly in storing materials.

This article has focused on new memory technologies for filling the gap between electronically and mechanically addressed memories—that is, on electronic memories in the range 10^6 to 10^9 or 10^{10} bits. There is another gap between mechanically accessed disk memories and manually accessed records, such as huge stacks of magnetic tapes. The Tetrabit, IBM 3850, and Unicon memory systems fill that gap by mechanically addressing 10^{12} bits. Thus there has been progress in making electronic what was mechanical and mechanical what was manual. However, we are still far from the ideal shoe-box device with 10^{12} bits accessible in nanoseconds, and still farther from the capacities of 10^{15} bits needed for many already well-defined applications. Although much can still be expected from LSI and magnetic recording techniques, these greater goals may require radically new approaches.

Many laboratories are looking into basic principles. Superconductive memories based on the Josephson effect may be able to operate in picoseconds on miniscule power. The boundaries within the walls of magnetic domains, exploited in the bubble lattice devices, are also used in a so-called cross-tie memory that may provide nonvolatile storage memories on LSI chips (29). Other possibilities being explored are storage by electrolytic deposition in an array of cells, novel sources of electrons and ions for beam devices and applications of biomolecular properties. Most of this research is in an early laboratory stage and the eventual system aims are not sharply defined. Very high speed and very low power memories rather than large capacity may well be the benefits of some of these approaches.

Any radical improvement in memory technology will ultimately greatly affect our way of life, as previous innovations have shown. The challenge is a fascinating one as it reaches into basic knowledge and demands imaginative invention.

Summary

Computers today use a hierarchy of large-capacity, relatively slow mechanically accessed memories in conjunction with fast electronically accessed memories of relatively small capacity. While the gap between these is spanned by ingenious organizations and programming, it would be highly desirable to fill it instead by some device of sufficient capacity and speed. Candidates for gap-filling memories include metal oxide semiconductor (MOS) random-access memories (RAM's) made by large-scale integration (LSI); charge-coupled devices; magnetic bubble devices based on cylindrical domains of magnetization; electron beam–addressed memories; and optical memories based on lasers, holography, and electrooptical effects. At present, the MOS RAM is the prime contender. Its natural evolution and the evolution of magnetic-recording techniques on which mass storage is based are likely to continue to shape the future as they have for more than a decade. On the other hand, radically new technologies, still at an early laboratory stage, are aimed at a more ideal solution than today's hierarchy.

References

1. E. W. Pugh, *IEEE Trans. Magn.* 7, 810 (1971).
2. A comprehensive review of digital storage systems appears in *Proc. IEEE* 63 (No. 8) (1975).
3. G. C. Feth, *IEEE Spectrum* 13, 37 (June 1976).
4. J. W. Forrester, *J. Appl. Phys.* 22, 44 (January 1951).
5. J. A. Rajchman, *RCA Rev.* 13, 183 (June 1952).
6. ———, *Sci. Am.* 217, 48 (July 1967).
6a. F. P. Brooks, Jr., *IEEE Trans. Magn.* 15 (No. 3) (September 1969).
7. R. Stone, *Science* 196, 1102 (1977).
8. J. D. Schmidt, *Solid State Design* (January 1965), pp. 21–25.
9. D. A. Hodges, *Proc. IEEE* 63, 1136 (1975).
10. R. A. Abbott, W. M. Regitz, J. A. Karp, *IEEE J. Solid State Circuits* SC-8, 299 (October 1973).
11. C. Koo, N. Kitagawa, D. Ogden, *Electronics* 49, 81 (May 1976); also commercial announcements of various manufacturers in 1976.
12. W. B. Sander, W. H. Sheperd, R. D. Schinelle, *ibid.* 49, 99 (August 1976).
13. A. N. Broers and M. Hatzakis, *Sci. Am.* 227, 34 (November 1972).
14. R. Bakish, Ed., *Proceedings of the Sixth International Conference on Electron and Ion Beam Science and Technology* (Electrochemical Society, Princeton, N.J., 1976).
15. W. S. Boyle and G. E. Smith, *Bell Syst. Tech. J.* 49, 587 (1970).
16. L. M. Terman and L. G. Heller, *IEEE J. Solid State Circuits* SC-11, 4 (February 1976).
17. A. M. Mohsen, M. F. Tompsett, E. N. Fuls, E. J. Zimanny, Jr., *ibid.*, p. 40.
18. W. F. Kosonocky, *ONR Contract Rep. Contract N00014-76-C-0371* (1976), pp. V-520 to V-534.
19. A. H. Bobeck, *Bell Syst. Tech. J.* 46, 1901 (1967).
20. ———, P. J. Bonyhard, J. E. Geyer, *Proc. IEEE* 63, 1176 (1975).
21. M. S. Cohen and H. Chang, *ibid.*, p. 1196.
22. P. J. Bonyhard and J. L. Smith, *IEEE Trans. Magn.* 12, 614 (1976).
23. G. S. Almasi, G. E. Keefe, Y. S. Lin, R. J. Hendel, R. F. McGovey, paper presented at the INTERMAG Conference, London, April 1975.
24. O. Voegli, B. A. Calhoun, E. L. Rosier, J. C. Slonczewski, *AIP Conf. Proc.* 24, 617 (1975).
25. W. C. Hughes, C. O. Lemmond, H. G. Parks, G. W. Ellis, G. E. Possin, R. H. Wilson, *Proc. IEEE* 63, 1230 (1975).
26. D. Chen and J. D. Zook, *ibid.*, p. 1207.
27. J. A. Rajchman, *J. Appl. Phys.* 41, 1376 (1970).
28. W. C. Stewart, R. S. Mezrich, L. S. Cosentino, E. M. Nagle, F. S. Wendt, R. D. Lohman, *RCA Rev.* 34, 3 (March 1973).
29. L. J. Schwee, H. R. Irons, W. E. Anderson, *IEEE Trans. Magn.* 23, 608 (1976).

Microelectronics: Lithographic Technologies Progress

Arthur L. Robinson

For reasons of economics and performance, it is desirable to miniaturize the microelectronic devices known as integrated circuits to the maximum extent possible. But diffraction of the ultraviolet light used in the photolithographic process that delineates the circuits limits the minimum size of their features. Researchers have speculated for several years that one day the use of beams of electrons to delineate circuit patterns would overcome the diffraction limitation in that the wavelength of the electrons would be only a fraction of an angstrom. While formidable problems have slowed the development of electron beam lithography, some recent advances indicate progress.

Groups led by Hwa Yu, Robert Dennard, and Philip Chang at IBM's Thomas J. Watson Research Center, Yorktown Heights, New York, working together, used a scanning electron beam apparatus controlled by a computer to fabricate an 8192-bit random access memory with an area of 1.8 square millimeters, by far the most complex and densely packed device yet made by this technique. At Bell Laboratories, Murray Hill, New Jersey, researchers under the direction of Donald Herriott have developed a quite different type of scanning electron beam system. This system now routinely produces high-quality masks that are then used in conventional photolithography to determine what locations on an integrated circuit chip are exposed to the ultraviolet light. Finally, research under way in several laboratories on a photolithographic process in which x-rays are used in place of ultraviolet light suggests that the high resolution inherent in the scanning electron beam technique may be combined with the economical aspects of photolithography by using electron beams to produce masks for short wavelength and hence largely undiffracted x-rays.

There is good reason for increasing the degree of miniaturization. Speed of operation is enhanced because electrons have smaller distances to travel. Circuits are more reliable when more components are on a common substrate because wiring between devices is reduced. And the cost per electric function (such as storing a binary bit of information or making a logic decision) decreases because the cost of producing a complex integrated circuit does not differ greatly from that of a simple one. Miniaturization of electronic circuits has already permitted the development of products including accurate long-range missiles and high-speed super computers as well as small inexpensive pocket calculators and digital wristwatches.

At present the transistors, diodes, capacitors, resistors, and interconnections that together comprise an integrated circuit are delineated in a series of steps, one of which is the contact printing process called photolithography. Ultraviolet light shining through a mask carrying a circuit pattern exposes the pattern in a layer of photosensitive organic polymer, or photoresist, on a silicon wafer. When the exposed photoresist is developed, either the exposed (positive resist) or the unexposed (negative resist) regions dissolve. The undissolved resist outlines those areas of the silicon to be subsequently either impregnated with impurity dopants that control its electrical properties or coated with a metallic conductor that interconnects the components of the circuit. Several masks and sequences of such processing operations are needed to make a complete circuit.

The mask itself consists of a glass substrate covered with a photographic emulsion or a harder material, such as chromium, in which the circuit pattern is produced photographically. Depending on the distance between the mask and the silicon wafer during the exposure step, diffraction of the 4000-angstrom ultraviolet light, variations in the wafer flatness, errors incurred in transferring the original pattern to the final mask, accumulations of dirt in the mask, and distortions produced when the mask is pressed against the wafer variously conspire to limit the minimum size of features (usually the width of the metal interconnections) to from 2 to 7 micrometers.

While the minimum size of features that can be exposed photolithographically approaches the resolution of optical microscopes, the same cannot be said for the relation between electron beam pattern generation and electron microscopes. High resolution transmission electron microscopes can resolve features as small as 2 or 3 angstroms and special scanning electron microscopes can resolve a few tens of angstroms, but electron beams cannot be used to delineate such small dimensions. The fundamental limit is given neither by the electron wavelength nor even by the diameter of the electron beam. Instead, electron scattering effects that cause the energetic (10,000 to 25,000 electron volts) electrons effectively to expose a wider area of the electron sensitive resist than that delineated by the incident beam constitute a practical limit to obtaining small dimensions. This scattering phenomenon can be especially troublesome when several features with small dimensions are so closely spaced that the scattering regions overlap, although features as small as 500 angstroms have been produced in the laboratory.

(Since this article first appeared, Alec Broers and his colleagues at IBM have fabricated square grid-like patterns with minimum dimensions of 80 angstroms. The researchers first covered their sample (a 100-angstrom thick layer of graphite) with a metal alloy of gold and palladium of about the same thickness, but used no resist. Instead, an electron beam 5 angstroms in diameter caused the build up of a thin, impervious layer of polymer on the alloy only where the beam struck the surface. The polymer formed from contaminants in the vacuum system used. Since the sample layer was only 100 angstroms thick, the backscattering of electrons, which blurred patterns made by conventional techniques, was drastically reduced.)

Researchers at IBM did not attempt to utilize the maximum achievable resolution of an electron beam system in fabricating their 8192-bit random access memory (Fig. 1). Instead, the investigators settled for demonstrating that the best available silicon processing technology was compatible with a tenfold increase in the density of components on a chip. They redesigned a circuit that scientists at IBM's Essex Junction, Vermont, facility had earlier fabricated using conventional technology, and scaled down the entire circuit so that the area was reduced by a factor of about 10 with a chip size of 1.1 by 1.6 millimeters and

The author is a member of the staff of the Research News Section of *Science*.

with minimum dimensions of about 1.25 micrometers.

Unlike photolithography, in which the entire pattern is exposed at once, scanning electron beam lithography is a serial process in which the electron beam writes the circuit pattern point by point. The scanning may be a raster scan akin to the sweeping of the electron beam in a television picture tube, except that the beam is turned on only when a position being scanned is to be exposed. In the so-called vector scan method, the beam goes directly to points that need exposing and does not waste time passing over points that will not be exposed.

In either case, a mask is not required because the beam is controlled electronically. For example, in vector scanning, a computer containing the circuit pattern in its memory can direct the beam to the starting coordinates of certain basic shapes, such as rectangles, into which the pattern has been broken up. These elements are then filled in by a rasterlike scan. The requirements on the system that deflects the beam of electrons on its way from the electron source to the target are quite stringent for this type of scanning, in part because the magnetic fields that deflect the electron beam produce eddy currents in metal parts of the system which, in turn, generate small magnetic fields. These fields can distort the image traced out by slightly deflecting the beam. Corrections are required for the interaction of the focusing fields and the fields produced by the deflection system.

The scanning system developed at IBM by Chang and his co-workers has a beam size that can be varied from 500 angstroms to 0.5 micrometer and a scanning field of 8000 beam positions in each direction. The minimum size of circuit features is usually 4 or 5 beam diameters. Thus fields with dimensions of 2000 line widths on a side can be exposed. The IBM researchers used a beam diameter of 0.25 micrometer with a concomitant field size of 2 millimeters. The electron optical system also has a longer focal length than scanning electron microscopes in order to permit the large deflections of the electron beam needed to write over a 2-millimeter distance while keeping the deflection angle small to avoid distortions.

Many observers think, however, that it will be very difficult to simultaneously increase the size of the field much further toward the 3-millimeter dimensions of the largest present-day integrated circuit chips and to maintain submicrometer resolution. Increasing the focal length also increases aberrations in the magnetic

Fig. 1. Scanning electron micrograph of a small portion of the 8192-bit memory array fabricated at IBM. The vertical, light-colored lines are aluminum interconnections about 2.2 micrometers wide. Each short cross line on the aluminum contacts one memory cell. The narrow horizontal lines also contact each memory cell and have widths from 1 to 1.5 micrometers. [Source: R. T. Miller, IBM]

lenses that focus the electron beam, thus reducing the electron beam current and the speed with which patterns can be written. One partial solution developed at Texas Instruments in Dallas by Gilbert Varnell and his co-workers is called dynamic focusing. As the deflection increases, the focusing of the electron beam is changed to account for the distortion. The Texas Instruments researchers are using a scanning electron beam system with dynamic focusing under the control of a computer to make a 256-bit random access memory with lateral dimensions reduced by a factor of 4 as compared with its photolithographically fabricated counterpart.

Integrated circuits are actually three-dimensional entities and must be reduced in size proportionately in all three dimensions, points out IBM's Yu. Therefore, even such aspects as the concentration of impurity dopants and operating voltages must be scaled accordingly when the size is decreased.

For example, the traditional method of incorporating doping impurities into semiconductors is to coat the semiconductor surface with the impurity or one of its compounds and cause it to diffuse a short distance into the surface by heating the semiconductor to a high temperature. However, it is difficult to control the depth to which the impurities penetrate when the depth becomes much less than 0.5 micrometer. A method now coming into commercial use and called ion implantation overcomes this problem. When ions of the desired impurity bombard the semiconductor surface, both the depth of penetration of the ions and the distribution within the semiconductor can be controlled to a certain extent by the energy of the incident ions and by the relative orientation of the ion beam and the crystal lattice of the semiconductor. The IBM scientists used this method of doping exclusively during the fabrication of the memory circuit to make doped regions as shallow as 0.2 micrometer.

Apart from the still-limited size of the scanning field, a principal difficulty with the serial scanning of electron beams is the relative slowness of the process. Chang has estimated that his system could expose a 2-millimeter square field having features 1 micrometer in width in less than 10 seconds. But a typical semiconductor wafer today has a diameter of 5 to 7.5 centimeters and thus hundreds of integrated circuit chips are made from one wafer. With photolithography, it takes about 10 seconds to expose an entire wafer through one mask. The exposure speed of the scanning electron system could be increased by increasing the electron beam current, although this solution has problems of distortion associated with it. Another solution is to find more sensitive electron resists. But even with these improvements, high speed scanning systems would remain to be developed.

(In another recent development, H. S. Yourke and E. V. Weber of IBM's Systems Products Division, Hopewell Junction, New York, and their associates announced the operation of a high-throughput scanning electron beam lithography system for semiconductor manufacture. The system uses a square, rather than round, electron beam that is the same size as the minimum dimension needed in the pattern generated. Thus, the system can be 25 times as fast as the system designed by Chang and his colleagues. At present, 22 wafers, each 5.5 centimeters in diameter, can be processed per hour. The minimum possible dimension is, however, not as small as that in Chang's apparatus, being about 2.5 microns.)

Mindful of these difficulties, researchers at Bell Laboratories have built a conservatively designed scanning electron beam system which they have named the electron beam exposure system (EBES). This system is used about half the time for fabrication of experimental devices, such as a 1024-bit random access memory whose lateral dimensions were halved as compared to its conventional counter-

part. The remainder of the time EBES produces masks for other laboratories at Bell to use in photolithographic production of integrated circuits or other devices. According to Herriott and Fabian Pease of Bell Laboratories, production of masks in one step by scanning electron beams is both faster and more economical than the older methods involving several pattern generation, reduction, inspection, and duplicate mask-making steps.

Equally important is the improved quality of the mask made by EBES as compared to ordinary masks, so that even when the EBES masks do not have smaller features than conventional ones, they have many fewer defects and consequently the yield (ratio of finished circuits that work to the total number manufactured) is increased.

(Recently, Bell Laboratories concluded licensing arrangements with two companies—Etec Corporation, Hayward, California, and the Extrion division of Varian Associates, Gloucester, Massachusetts—allowing them to market the EBES system. Both Etec and Extrion say they will begin shipping units during 1977.)

Overall, EBES differs substantially from Chang's system at IBM. Rather than a scanning field whose width is based on a beam having 8000 positions, the Bell system has a scanning field only 128 micrometers wide based on a 0.5-micrometer beam having 256 positions. The immediate consequence is that the distortions due to aberrations and other sources are for the most part absent, so that complex compensating mechanisms do not need to be devised. Also EBES uses a raster scan, which minimizes data processing and hence can be quite fast.

For writing over areas larger than this small field size, a continuously moving mechanical table under the control of a laser interferometer was developed. The interferometer monitors the position of the table (and the wafer or mask sitting on it) and feeds back the position information to an error control system that repositions the electron beam if the table is not where it is supposed to be. (The use of a laser interferometer in conjunction with a scanning electron beam system was pioneered by O. Cahen, R. Sigelle, and J. Trotel of Thomson-CSF, a French company. But the French workers did not have a continuously moving table.) With this combined moving table and scanning electron beam system, areas up to 100 millimeters on a side can be exposed.

For example, a circuit pattern to be repeated on each chip of a wafer containing several integrated circuit chips is decomposed into strips 128 micrometers wide. The core memory of the controlling computer is loaded with the first 128-micrometer-wide segment. This portion of the pattern is successively exposed on each of the chips on the wafer as the table moves the wafer under the electron beam. Then the table is repositioned at the starting point, the second 128-micrometer strip is written in the same way as the first. This cycle is repeated until each segment has been written on every chip on the wafer. With this method and with electron resists (developed by L. F. Thompson, M. J. Bowden, and their co-workers) that are about 100 times as sensitive as the polymethyl methacrylate that is commonly used in electron beam lithography, the Bell scientists can achieve an exposure speed of 1 square centimeter per minute.

Fig. 2. Scanning electron micrograph of half-micron wide patterns formed in photoresist by exposure to x-rays. The mask used to define the pattern was made in a computer-controlled electron beam fabrication system. [Source: R. T. Miller, IBM]

In order to accurately position the table before each strip is exposed (and also to obtain an occasional check on the position of the electron beam during the exposure), registration or fiducial markers are used. These are raised areas on the wafer whose position can be monitored by the observation of electrons that are scattered when the electron beam strikes the markers. A similar scheme is used at both IBM and Texas Instruments and is essential in order to ensure that successive exposures (corresponding to the several masks needed to make a complete circuit) of a circuit pattern are properly aligned. A set of three such markers can be used to adjust the size and orthogonality of the x and y axes as well as the position and orientation of the pattern. Alignments to within 0.1 micrometer are possible with this method. The IBM researchers have also used such registration markers to "stitch together" adjoining fields—like tiles on a floor—thus effectively enlarging the field size.

The capital cost of scanning electron beam systems, researchers point out, may amount to $1 million. Thus, it is obvious why manufacturers would hesitate to invest in the several scanning systems required to make production quantities of integrated circuits if there were any alternative. And there are at least two alternatives. For either of these, the expensive scanning electron beam machines would be restricted to producing masks (pattern generation) with the fine features needed. Production of actual circuits (pattern replication) would be analogous to the present photolithography with electrons or x-rays replacing ultraviolet light. The first alternative involves projecting electrons through a mask. Although various such devices have been tested, problems associated with both distortions and registration have yet to be satisfactorily solved. The second alternative is a system that involves projecting x-rays through a mask.

Henry Smith and his associates at the Massachusetts Institute of Technology's Lincoln Laboratories in Lexington described an x-ray lithographic technique in 1972 which scientists say is primarily responsible for the current interest in x-rays. The Lincoln Laboratory investigators made masks (by conventional techniques) of thin layers of gold on silicon substrates. For the 8.3-angstrom x-ray radiation used, gold is absorbing and silicon is transmitting. Some researchers are now using Mylar as a substrate, because it is more convenient to work with and is more transparent at some x-ray wavelengths than silicon.

For example, Daniel Maydan and his colleagues at Bell Laboratories use Mylar substrates in conjunction with short wavelength 4.6-angstrom x-rays. The shorter wavelength x-rays are also attenuated less than the longer wavelength x-rays by beryllium, so that this material can be used as a window to separate the evacuated x-ray source from the target chamber for rapid sample changes.

Advantages of X-rays

Because of the short wavelength of the x-rays, diffraction effects are not a problem, and unlike electrons, x-rays are not readily scattered, so that the resolution of patterns made with x-ray lithography is limited mainly by that of the mask, and by certain distortions that arise because

the mask does not sit in contact with the wafer beneath it. Another advantage of x-rays was recently described by Ralph Feder, Eberhard Spiller, and John Topalian of IBM. These investigators found that features with high aspect ratios—that is, whose depth is much greater than their width—can be exposed in resists by x-rays. With electron beams, aspect ratios of at best 3 : 1 or 4 : 1 can be achieved, as compared with 15 : 1 or more with x-rays. This effect might enable, for example, very narrow metallic interconnections to carry high currents, because the cross-sectional area could remain high (Fig. 2). Finally, scratches or dust on the mask are not a problem for x-rays because of their short wavelength and their penetrating power, whereas they are a considerable problem in photolithography.

A major difficulty with x-rays is that it takes a high x-ray flux to expose many of the existing resists. With available x-ray sources, exposure times until recently were more than 100 times those of conventional photolithography. Robert Brault of Hughes Research Laboratories, Malibu, California, has developed a metal acrylate resist that is some 50 times more sensitive than polymethyl methacrylate. The addition of metals, such as barium, which absorb x-rays efficiently means that the resist has a higher absorption coefficient for x-rays than do organic resists containing only light atomic weight elements.

A second approach is to raise the x-ray flux, but a fundamental limitation is the melting of the metal target from which x-rays are produced by bombarding it with electrons. One partial solution is the use of rotating metal targets in the x-ray tube which can increase its output by about 20 times as compared with tubes with stationary targets because the entire target is not heated at once. Another tactic that has been tried is to use the high-intensity x-rays from synchrotron radiation sources, such as SPEAR electron-positron storage ring at the Stanford Linear Accelerator Center (*Science*, 12 December 1975, p. 1074). Although not considered so absurd an idea as to be unthinkable, the economics of such a scheme would seem to be questionable.

Registration is a second major problem with the x-ray technique. At present a completely satisfactory method for aligning a mask with registration marks has not been developed, and hence use of x-rays is now restricted to those devices that can be made with only one mask. The optical techniques used with photolithography can also be used with x-rays, but may be of limited value when circuits having features with dimensions of 1 micrometer or less are needed. John McCoy and Paul Sullivan of Hughes are working on an automatic alignment system that electronically detects the position of x-rays passing through transparent regions of the mask. Piezoelectric transducers controlled by a feedback system connected to the detectors would physically position the mask to within 0.1 micrometer, comparable to the alignment accuracy of scanning electron beam systems. However, problems with the electronics have prevented achieving this level of performance.

At present, there is no alternative to conventional photolithography that does not have unsolved problems. Even when perfected, scanning electron beam lithography may turn out to be too expensive for direct fabrication of integrated circuits. Although emphasizing that it is still too early to make a sure prediction, investigators manage nonetheless to convey that the best hope for the future may lie in the combination of scanning electron beam systems to generate masks with circuit fabrication by a projection technique, such as x-ray lithography.

Additional Reading

IEEE Transactions on Electron Devices, ED-22 (July 1975). The entire issue is devoted to lithographic techniques.

Mathematics and Computer Science: Coping with Finiteness

Advances in our ability to compute are bringing us substantially closer to ultimate limitations.

Donald E. Knuth

A well-known book entitled *One, Two, Three, . . . Infinity* was published by Gamov about 30 years ago (*1*), and he began by telling a story about two Hungarian noblemen. It seems that the two gentlemen were out riding, and one suggested to the other that they play a game: Who can name the largest number. "Good," said the second man, "you go first." After several minutes of intense concentration, the first nobleman announced the largest number he could think of: "Three." Now it was the other man's turn, and he thought furiously, but after about a quarter of an hour he gave up. "You win," he said.

In this article I will try to assess how much further we have come, by discussing how well we can now deal with large quantities. Although we have certainly narrowed the gap between three and infinity, recent results indicate that we will never actually be able to go very far in practice. My purpose is to explore relationships between the finite and the infinite, in the light of these developments.

Some Large Finite Numbers

Since the time of Greek philosophy, men have prided themselves on their ability to understand something about infinity; and it has become traditional in some circles to regard finite things as essentially trivial, too limited to be of any interest. It is hard to debunk such a notion, since there are no accepted standards for demonstrating that something is interesting, especially when something finite is compared with something transcendent. Yet I believe that the climate of thought is changing, since finite processes are proving to be such fascinating objects of study.

In the first place, it is important to understand that finite numbers can be extremely large. Let us start with some very familiar and fairly small numbers: the value of xn is $x + x + \cdots + x$, added n times. Similarly we can define a number I shall write as $x\uparrow n$, which means $xx\cdots x$ multiplied n times. For example, $10\uparrow 10 = 10\cdot10\cdot10\cdot10\cdot10\cdot10\cdot10\cdot10\cdot10\cdot10 = 10{,}000{,}000{,}000$ is 10 billion; this is usually written 10^{10}, but it will be clear in a minute why I prefer to use an upward arrow. In fact, the next step uses two arrows

The author is professor of computer science at Stanford University, Stanford, California 94305.

$$x\uparrow\uparrow n = x\uparrow(x\uparrow(\cdots\uparrow x)\cdots))$$

where we take powers n times. For example

$$10\uparrow\uparrow 10 = 10^{10^{10^{10^{10^{10^{10^{10^{10^{10}}}}}}}}}$$

$$= 1 \text{ followed by } 10^{10^{10^{10^{10^{10^{10^{10^{10^{10}}}}}}}}} \text{ zeros}$$

This is a pretty big number; at least, if a monkey sits at a typewriter and types at random, the average number of trials before he types perfectly the entire text of Shakespeare's *Hamlet* would be much, much less than this: it is merely a 1 followed by about 40,000 zeros. The general rule is

$$\overbrace{x\uparrow\uparrow\cdots\uparrow n}^{k \text{ arrows}} = \underbrace{x\overbrace{\uparrow\cdots\uparrow}^{k-1}(x\overbrace{\uparrow\cdots\uparrow}^{k-1}(\cdots\overbrace{\uparrow\cdots\uparrow}^{k-1}x)\cdots))}_{n \text{ times}}$$

Thus, one arrow is defined in terms of none, two in terms of one, three in terms of two, and so on.

In order to see how these arrow functions behave, let us look at a very small example

$$10\uparrow\uparrow\uparrow\uparrow 3$$

This is equal to

$$10\uparrow\uparrow\uparrow(10\uparrow\uparrow\uparrow 10)$$

so we should first evaluate $10\uparrow\uparrow\uparrow 10$. This is

$$10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow 10)))))))))$$

and that is

$$10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow 10^{10^{10^{10^{10^{10^{10^{10^{10^{10}}}}}}}}})))))))))$$

$$= 10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow 10^{10^{10^{10^{10^{10^{10^{10^{10^{10}}}}}}}})))))))))$$

where the stack of 10's is $10\uparrow\uparrow 10$ levels tall. We take the huge number at the right of this formula, which I cannot even write down without using the arrow notation, and repeat the double-arrow operation, getting an even huger number, and then we must do the same thing again and again. Let us call the final result $\mathcal{H}$. (it is such an immense number, we cannot use just an ordinary letter for it.)

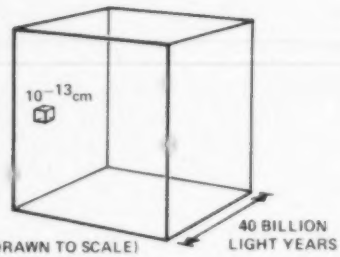


Fig. 1. The known universe fits inside this box.

Of course we are not done yet, we have only evaluated $10\uparrow\uparrow\uparrow 10$; to complete the job we need to stick this gigantic number into the formula for $10\uparrow\uparrow\uparrow\uparrow 3$, namely

$$10\uparrow\uparrow\uparrow\uparrow 3 = 10\uparrow\uparrow\uparrow\mathcal{H}$$

$$= \underbrace{10\uparrow\uparrow(10\uparrow\uparrow(10\uparrow\uparrow\cdots\uparrow\uparrow(10\uparrow\uparrow 10)\cdots))}_{\mathcal{H} \text{ times}}$$

The three dots " ⋯ " here suppress a lot of detail—maybe I should have used four dots. At any rate it seems to me that the magnitude of this number $10\uparrow\uparrow\uparrow\uparrow 3$ is so large as to be beyond human comprehension.

On the other hand, it is very small as finite numbers go. We might have used $\mathcal{H}$ arrows instead of just four, but even that would not get us much further—almost all finite numbers are larger than this. I think this example helps open our eyes to the fact that some numbers are very large even if they are merely finite. Thus, mathematicians who stick mostly to working with finite numbers are not really limiting themselves too severely.

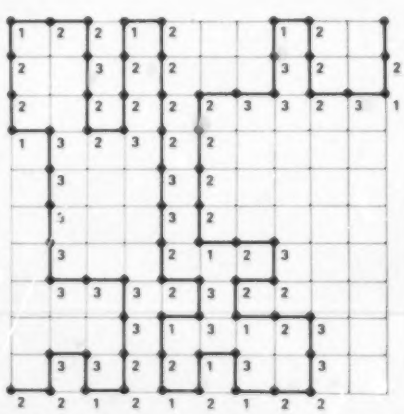


Fig. 2. A "random" path from the lower left corner to the upper right corner of a 10 × 10 grid.

Realistic Numbers

This discussion has set the stage for the next point I want to make, namely that our total resources are not actually very large. Let us try to see how big the known universe is. Archimedes began such an investigation many years ago, in his famous discussion of the number of grains of sand that would completely fill the earth and sky; he did not have the benefit of modern astronomy, but his estimate was qualitatively the same as what we would say today. The distance to the farthest observable galaxies is thought to be at most about 10 billion light years. On the other hand, the fundamental nucleons that make up matter are about 10^{-12} centimeter in diameter. In order to get a generous upper bound on the size of the universe, let us imagine a cube that is 40 billion light years on each side, and fill it with tiny cubes that are smaller than protons and neutrons, say 10^{-13} cm on each side (see Fig. 1). The total number of little cubes comes to less than 10^{125}. We might say that this is an "astronomically large" number, but actually it has only 125 digits.

Instead of talking only about large numbers of objects, let us also consider the time dimension. Here the numbers are much smaller; for example, if we take as a unit the amount of time that light rays take to travel 10^{-13} cm, the total number of time units since the dawn of the universe is only one fourth the number of little cubes along a single edge of the big cube in Fig. 1, assuming that the universe is 10 billion years old.

Coming down to earth, it is instructive to consider typical transportation speeds.

Snail	0.006	mile/hour
Man walking	4	mile/hour
U.S. automobile	55	mile/hour
Jet plane	600	mile/hour
Supersonic jet	1200	mile/hour

I would never think of walking from California to Boston, but the plane flight is only 150 times faster. Compare this to the situation with respect to the following computation speeds, given 10-digit numbers.

Man (pencil and paper)	0.2/sec
Man (abacus)	1/sec
Mechanical calculator	4/sec
Medium-speed computer	200,000/sec
Fast computer	200,000,000/sec

A medium-fast computer can add 1 million times faster than we can, and the fastest machines are 1000 times faster yet. Such a ratio of speeds is unprecedented in history: consider how much a mere factor of 10 in speed, provided by

the automobile, has changed our lives, and note that computers have increased our calculation speeds by six orders of magnitude; that is more than the ratio of the fastest airplane velocity to a snail's pace.

I do not mean to claim that computers do everything a million times faster than people can; mere mortals like us can do some things much better. For example, you and I can even recognize the face of a friend who has recently grown a moustache; and for tasks like ling, a computer may be only ten or so times faster than a good secretary. But when it comes to arithmetic, computers appear to be almost infinitely fast compared with people.

As a result, we have begun to think about computational problems that used to be unthinkable. Our appetite for calculation has caused us to deal with finite numbers much larger than those we considered before, and this has opened up a rich vein of challenging problems, just as exciting as the problems about infinity which have inspired mathematicians for so many centuries.

Of course, computers are not infinitely fast, and our expectations have become inflated even faster than our computational capabilities. We are forced to realize that there are limits beyond which we cannot go. The numbers we can deal with are not only finite, they are very finite, and we do not have the time or space to solve certain problems even with the aid of the fastest computers. Thus, the theme of this article is coping with finiteness: What useful things can we say about these finite limitations? How have people learned to deal with the situation?

Advances in Technology and Techniques

During the last 15 years computer designers have made computing machines about 1000 times faster. Mathematicians and computer scientists have also discovered a variety of new techniques, by which many problems can now be solved enormously faster than they could before. I will present several examples of this; the first one, which is somehow symbolic of our advances in arithmetic ability, is the following factorization of a very large number, completed in 1970 by Morrison and Brillhart (*2*)

340,282,366,920,
938,463,463,374,607,431,768,211,457
= 5,704,689,200,685,129,054,721 ×
59,649,589,127,497,217

The point, of course, is not simply to compute the exact 39-digit product of these two large numbers; that is trivial and takes only a few millionths of a second. The problem is to start with the big 39-digit number and to discover its prime factors. (The big number is $2^{128} + 1$, and its factors are of use, for example, in the design of codes of a type used for space communications.) The number of microseconds per year is only 31,556,952,000,000, a 14-digit number, so even if we could test 1 million factors every second it would take about 2000 years to discover the smaller factor. The factorization actually took about 1½ hours of computer time; it was achieved by a combination of sophisticated methods representing a culmination of mathematical developments that began about 160 years earlier.

Latin Squares

Now let us look at another kind of example. Here is a so-called latin square of order 8, an arrangement of eight numbers in eight rows and eight columns so that each number appears in each row and each column.

1	2	3	4	5	6	7	8
2	1	4	3	6	5	8	7
3	4	1	2	7	8	5	6
4	3	2	1	8	7	6	5
5	6	7	8	1	2	3	4
6	5	8	7	2	1	4	3
7	8	5	6	3	4	1	2
8	7	6	5	4	3	2	1

On top of this square we can overlay another latin square of order 8, using italic numbers; again, there is one italic digit of every kind in every row and in every column.

1*1*	2*2*	3*3*	4*4*	5*5*	6*6*	7*7*	8*8*
2*3*	1*4*	4*1*	3*2*	6*7*	5*8*	8*5*	7*6*
3*5*	4*6*	1*7*	2*8*	7*1*	8*2*	5*3*	6*4*
4*7*	3*8*	2*5*	1*6*	8*3*	7*4*	6*1*	5*2*
5*4*	6*3*	7*2*	8*1*	1*8*	2*7*	3*6*	4*5*
6*2*	5*1*	8*4*	7*3*	2*6*	1*5*	4*8*	3*7*
7*8*	8*7*	5*6*	6*5*	3*4*	4*3*	1*2*	2*1*
8*6*	7*5*	6*8*	5*7*	4*2*	3*1*	2*4*	1*3*

These two latin squares are called orthogonal, since the superposition shows that every pair of roman and italic numbers occurs exactly once. Thus we have roman 1 with italic *1* (in the upper left corner), roman 1 with italic *2* (near the lower right corner), and so on; all 8 × 8 possibilities appear. Latin squares and orthogonal latin squares are commonly used in the design of statistical experiments and for such things as crop rotation.

The great 18th-century mathematician Euler showed how to construct pairs of orthogonal latin squares of all sizes except for order 2, 6, 10, 14, 18, and so on, and he stated his belief that orthogonal latin squares of these missing orders do not exist (*3*). It is easy to verify this for order 2; and in 1900, an exhaustive analysis by a French mathematician (*4*) showed that orthogonal latin squares of order 6 are indeed impossible. About 20 years later, an American mathematician (*5*) published a proof that Euler was right in the remaining cases 10, 14, 18, . . . ; but unfortunately his "proof" had a serious flaw so the question was still not settled. Finally computers were invented, and an attempt was made to test Euler's conjecture in the smallest remaining case, order 10.

In 1952, a group of mathematicians at the University of California at Los Angeles (UCLA) decided to see if there were any latin squares orthogonal to the following one of order 10.

0	1	2	3	4	5	6	7	8	9
1	8	3	2	5	4	7	6	9	0
2	9	5	6	3	0	8	4	7	1
3	7	0	9	8	6	1	5	2	4
4	6	7	5	2	9	0	8	1	3
5	0	9	4	7	8	3	1	6	2
6	5	4	7	1	3	2	9	0	8
7	4	1	8	0	2	9	3	5	6
8	3	6	0	9	1	5	2	4	7
9	2	8	1	6	7	4	0	3	5

This particular square was selected more or less at random, using a procedure analogous to one discussed in the next example below; the probability that the above latin square will be generated (*6*) turns out to be about 10^{-26}, so I imagine that there are extremely many 10 × 10 latin squares, something like 10^{26} at least. However, the computer at UCLA ran for many hours trying to find an orthogonal mate for this square; finally, having produced no answers, it was shut off (*7*). This was consistent with Euler's conjecture that no mates exist, but the investigators realized that several hundred more years of calculation would be required to show this exhaustively—and then they would have to try to find mates for the other 10^{26} or so initial squares.

The method used in this experiment was to look for a mate by filling in the entries row by row, one entry at a time in all possible ways, without violating the definition of orthogonal latin squares. Furthermore, they used the fact that the leftmost column of the orthogonal mate can be assumed to contain the digits 0 to 9 in order. Five years later Parker (*8*) discovered a far better way to look for orthogonal mates. His idea was to find all ways to put ten 0's into an orthogonal mate for a particular square; this means finding one entry in each row and each column so that no two entries contain the same digit. This is a much easier problem, and it turned out that there were roughly 100 ways to do it, using any cell

in the first column. The remaining problem is to combine a solution for the 0's with a solution for the 1's and a solution for the 2's, and so forth, and again this is comparatively simple. Parker was able to deduce that there is exactly one latin square orthogonal to the one studied at UCLA, namely the italic digits in the following array.

00	*12*	*28*	*35*	*49*	*54*	*67*	*73*	*86*	*91*
11	*87*	*34*	*29*	*53*	*46*	*75*	*60*	*92*	*08*
22	*95*	*56*	*64*	*38*	*07*	*80*	*41*	*79*	*13*
33	*76*	*09*	*90*	*84*	*65*	*18*	*52*	*21*	*47*
44	*68*	*71*	*57*	*25*	*93*	*06*	*89*	*10*	*32*
55	*01*	*97*	*48*	*70*	*82*	*39*	*14*	*63*	*26*
66	*59*	*40*	*72*	*17*	*31*	*23*	*98*	*04*	*85*
77	*43*	*15*	*81*	*02*	*20*	*94*	*36*	*58*	*69*
88	*30*	*62*	*03*	*96*	*19*	*51*	*27*	*45*	*74*
99	*24*	*83*	*16*	*61*	*78*	*42*	*05*	*37*	*50*

And the total time for his program to be completed, on a slow computer in 1959, was less than 1 minute.

This example, together with the previous example about factoring, illustrates an important point: we should never expect that the first way we try to do something on a computer is the best way. Good programming is much more subtle than that; chances are that an expert can find a method that will go considerably faster than that of a novice, especially in combinatorial problems where there have been significant advances in techniques during recent years. By analyzing Parker's method statistically, I estimate that his approach runs about 100 billion times faster than the original method used by the extremely competent mathematicians who studied this problem at UCLA; that is 11 orders of magnitude faster, because of a better idea.

By now many sets of orthogonal latin squares of order 10 have been found, and orthogonal pairs are known to exist for all orders greater than 6. But computers were of little help in discovering these facts; the constructions were discovered by hand (by Parker himself in many cases), generalizing from patterns observed in the smaller cases (*9*). For order 14 the problem is so much larger that even Parker's method would no longer be fast enough to search for all orthogonal mates by computer. This illustrates another point about combinatorial problems: the computation time often increases greatly when the size of the input to the problem has gone up only slightly.

Counting the Paths on a Grid

The next examples are all based on a single diagram, namely a grid of 100 squares; it is the diagram we would obtain if we drew boxes around the elements of a 10 × 10 latin square. (Incidentally, there are many possible examples that illustrate the points I wish to make, so it was necessary for me to find some way to narrow down the selection. Since a 10 × 10 array fits nicely on a page, I have decided to stick mostly to examples that are based somehow on this one diagram.)

First let us consider how many ways there are to go along the lines of such a grid from the lower left corner to the upper right corner, without touching the same point twice. Problems like this have been studied by chemists and physicists concerned with the behavior of large molecules (*10*); it seems to be a difficult problem, and no way is known to calculate the exact number of such paths in a reasonable amount of time. However, it is possible to obtain approximate solutions which are correct with high probability.

The idea is to construct a "random" path from the starting point to the finishing point. First we must go up or to the right; by flipping a coin or rolling some dice we might decide to go right. Again there are two choices, and half the time we will go up. From here there are three possibilities, and we may choose from these at random, say to the right. And so on. Figure 2 shows the first random path I generated in this way. At each choice point of Fig. 2, I have written the number of alternatives present when the path got that far. For example, the 1's at the edges mean that there is only one way to go, since the other way either is already occupied or leads into a trap.

The probability that this particular path would be obtained by such a random procedure is the product of all the individual probabilities at each choice point, namely

$$\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{3}\cdot\cdots\cdot\frac{1}{3}\cdot\frac{1}{1}\cdot\frac{1}{2}$$

$$= 2^{-343^{-24}}$$

$$= 1/4{,}852{,}102{,}490{,}441{,}335{,}701{,}504$$

about one chance in 5×10^{21}. So I am pretty sure that you have never seen this particular path before, and I doubt if I will ever generate it again.

In a similar vein, it is interesting to note that the great Mozart wrote a considerable amount of music that has never yet been performed. In one of his more playful moments, he specified 11 possibilities for each of the 16 bars of a waltz (*11*); the idea was that people from the audience should roll dice 16 times, obtaining a sequence of 16 numbers between 2 and 12 inclusive, and the performers would play the 16 bars corresponding to these respective rolls. The total number of ways to play Mozart's dice waltz is $2 \times 11^{11} = 759{,}499{,}667$, 966,482 (*12*); so it is safe to say that fewer than one out of every million of Mozart's melodies will ever be heard by human ear.

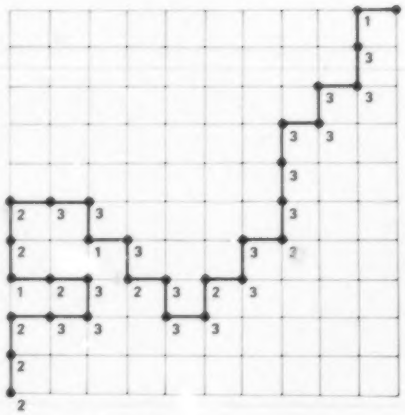


Fig. 3. A second path, which would be obtained with probability $\approx 3 \times 10^{-12}$.

Actually I have a phonograph record that contains 36 randomly selected waltzes from Mozart's scheme (*13*), and after hearing the fifth one or so I began to feel that the rest all sounded about the same. We might suspect that a similar thing will happen in this random path problem; all random paths from lower left to upper right might tend to look approximately like the first few. Figure 3 shows the second path I generated by making random choices; note that this one has quite a different character, and the strange thing is that the probability of obtaining it is more than ten orders of magnitude larger than we saw before. But still the probability is "negligibly small."

The third path I generated in this way decided to get into a corner and to hug the edge. The fourth one had its own twist; and the fifth was reminiscent of the first. These paths are shown in Fig. 4; of course I am displaying here each path exactly as I obtained it, not suppressing any that were uninteresting or unexpected, because the experiment must be unbiased.

The difference between this game and Mozart's dice music is that we know of no way to generate a truly random path, in the sense that each path should occur with the same probability. Although we have seen that each path occurs with extremely small probability, virtually zero, the actual probabilities differ from each other by many orders of magnitude.

If we want to estimate the total number of possible paths, solely on the basis of these data, a theorem of statistics tells us that the best estimate is obtained by using the average value of the reciprocals of the probabilities observed. Thus,

although three of these five paths had probabilities around 10^{-11}, suggesting that there are about 10^{11} possible paths, the much lower probabilities in the other two cases imply that it is much better to guess that there are about 10^{22} paths in all.

Based on the five experiments I have described, the best estimate of the average length of path will be about 70; and the best estimate of the chance that the point in the middle occurs somewhere on the path is that it almost always occurs, even though three-fifths of the experiments said the opposite. When large numbers like this are involved, we get into paradoxical situations, where the rules of statistics tell us that the best estimates are made by throwing away most of our data.

As you might expect, five experiments are not enough to determine the answers reliably. But by using a computer to generate several thousand random paths in the same way, I am fairly confident that the total number of possible paths from lower left to upper right is $(1.6 \pm 0.3) \times 10^{24}$, and that the average length of path is 92 ± 5. Conflicting evidence was obtained about the chance of hitting the center, but it seems that 81 ± 10 percent of all paths do hit the center point. Of course, I have only generated an extremely small fraction of these paths, so I cannot really be sure; perhaps nobody will ever know the true answer.

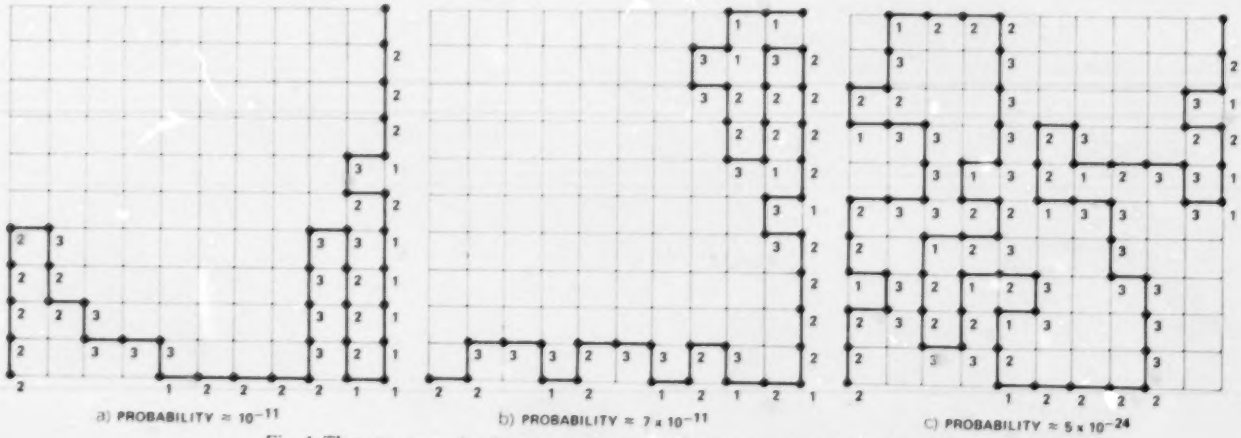
The Shortest Paths

For the next examples we will add weights to the lines in the grid. The basic diagram is shown in Fig. 5, where a random digit has been placed beside each line; these digits may be thought of as the lengths of roads between adjacent points of intersection. Thus, there are three roads of length 4 on the bottom line, and the upper part of the diagram contains three adjacent roads of length 0. Actually I must admit that the sequences of numbers are not completely arbitrary; for example, the reader might recognize 1.414213562... in the top line as the square root of 2, and π appears down the second column. For our purposes these digits will be random enough.

The first problem we might ask about such a network of roads is: What is the shortest route from the lower left corner to the upper right corner? We have estimated that there are some 10^{24} possible paths, and we might want to know which of these is shortest, using the given lengths.

Fortunately we do not have to try all possible paths to find the shortest; there is a simple method due to Dijkstra (*14*) which can be used to solve this problem by hand in less than half an hour. The answer (see Fig. 6) is a curious sort of path, which might very well be missed if one does not use a systematic method; it is the only way to go from southwest to northeast in a path of length 43.

The idea underlying Dijkstra's method is rather simple. Suppose that at some stage we have found all positions at distance 20 or less, say, from the southwest corner. By looking at the roads connecting these points to the others it will be easy to see which points will be at distance 21, and so on. You can imagine



a) PROBABILITY $\approx 10^{-11}$ b) PROBABILITY $\approx 7 \times 10^{-11}$ c) PROBABILITY $\approx 5 \times 10^{-24}$

Fig. 4. Three more randomly generated paths, with their associated probabilities.

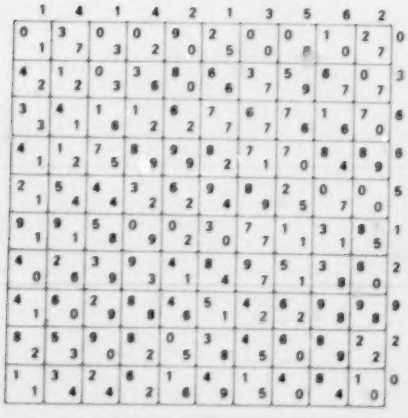


Fig. 5 (left). Network to be used in subsequent examples, based on 20 mathematical constants.

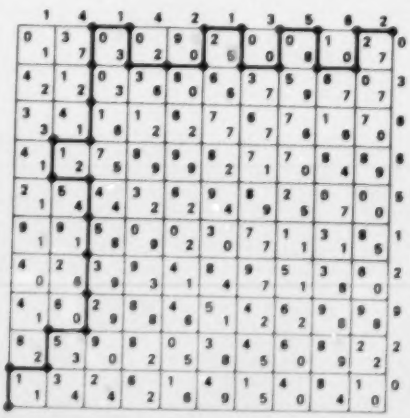


Fig. 6 (middle). The shortest route from lower left to upper right in this network of roads.

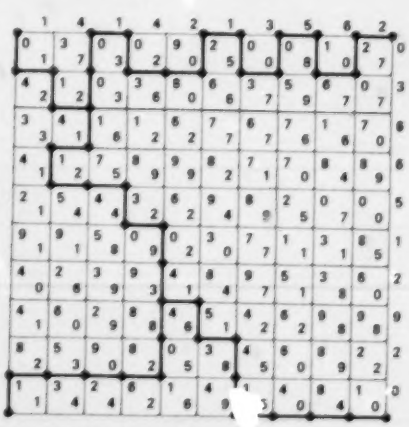


Fig. 7 (right). A shortest way to connect the four corners.

a fluid spreading over the diagram at the rate of one unit of length per minute.

Connecting Points in a Network

The next problem is somewhat harder. Suppose we want to construct electrical connections between all four of the corner points in Fig. 5: What is the shortest electrical network joining these four points, sticking to the lines and distances shown? Such a network is usually called a Steiner tree (*15*), and Fig. 7 shows the shortest possible one.

The number of possible Steiner trees connecting the four corners is much larger than the number of paths, but still I am sure that this is the shortest. In this case I do not know how to compute the shortest by hand, but a properly programmed computer can do it in a few seconds.

We say that we have a "good" algorithm for some problem if the time to solve it increases only as a polynomial in the size of the inputs; in other words, if doubling the size of the problem increases the solution time by at most a constant factor. There is a good algorithm to find Steiner trees connecting up to five points; it takes roughly n^3 steps, where n is the total number of points in the network of roads (*16*). But if we want to connect larger numbers of points by Steiner trees, the computation rapidly gets larger; and when the number of points to be connected is, say, as large as $n/10$, no good algorithm is known.

On the other hand, when our job is to find the shortest way to connect up all n of the points in the network, a good algorithm is available, again one that is so good it can be performed by hand in half an hour.

A minimal connection of all points in a network is called a spanning tree, and in the particular network we are considering it is possible to prove that the number of possible spanning trees is really huge, more than 4×10^{52}; in fact, the exact number (*17*) is 40,325,021,721,404,118,513,276,859,513,497,679,249,183,623,593,590,784

Yet we can find the best one, in a remarkably easy way discovered by Kruskal (*18*): simply consider all the lines one by one in order of increasing length, starting with the shortest one, then the next shortest, and so on. In case of ties between lines of the same length, use either order. The rule is to include each line in the spanning tree if and only if it connects at least two points that are not connected by a path of previously selected lines. This is called the greedy algorithm because it is based on the idea of trying the best conceivable possibilities first. Such a policy does not always solve a combinatorial problem—we know that greed does not always pay off in the long run—but in the case of spanning trees the idea works perfectly (see Fig. 8).

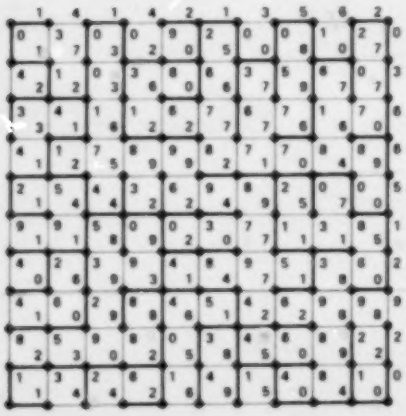
Maximum Matching

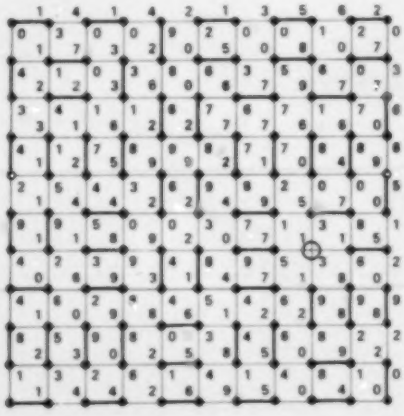
Another problem on this network for which a good algorithm is available is to choose 60 of the lines with the maximum possible sum, no two overlapping. We may think now of the points as people, instead of as cities, and the numbers now measure the amount of happiness generated between one person and his or her neighbor. The idea is to pair off the people so as to get the maximum total happiness. If men and women alternate in the diagram, with men at the corners, there will be 61 men and 60 women in all, so one man will have no partner; he makes a personal sacrifice for the greater good of the group as a whole. There are exactly 1,801,272,981,919,008 ways to do such a pairing, according to a mathematical theory worked out to solve a physical problem about crystals (*19*); Fig. 9 exhibits the best one.

It turns out that the circled man is the best to omit, and the others should pair up in this way. Once again we are able to find the optimum solution in 1 or 2 seconds on a computer if we use a suitable algorithm, even though the number of possible arrangements is far too large to examine exhaustively. In this case the algorithm is somewhat more subtle than the ones I have discussed earlier, but it is based on simple ideas. First we add a "dummy" woman who will be paired with the man who gets no real woman. The happiness rating is 0 between the dummy woman and every man. Then if we add or subtract some number from all the happiness ratings touching any particular person, the solution to the problem does not change. A clever way of adjusting these scores can be used so that all 61 of the ratings for the couples matched here are 9, and all the other ratings are 9 or less (*20*).

An Apparently Harder Problem

From these examples, one might get the idea that a good algorithm can be found for virtually any combinatorial problem. Unfortunately this does not appear to be true, although I did want to demonstrate that considerable progress has been made toward finding good methods. The next problem seems to be much harder: What is the shortest path from the lower left corner to the upper right corner that passes through all 121 points of the grid exactly once?





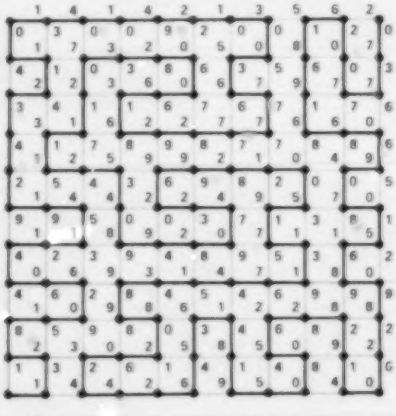


Fig. 8 (left). A minimum spanning tree. Fig. 9 (middle). The best choice of 60 nonoverlapping lines in the diagram. Fig. 10 (right). A shortest path from lower left to upper right, touching each point just once.

This is called the traveling salesman problem, because we might think of a salesman who wants to visit each city with minimum travel time. The problem arises frequently in industry—for example, when the goal is to find the best order in which to do n jobs, based on the costs of changing from one job to another. But it has resisted all attacks; we know how to solve medium-sized problems, but the algorithms are not good in the technical sense since the running time goes up rapidly on large cases.

The traveling salesman's path shown in Fig. 10 is as short as possible, and it required several minutes of computer time to verify the fact. To my knowledge, this is the largest network for which the traveling salesman problem has yet been solved exactly. I used a method suggested in 1971 by Held and Karp (*21*), based on a combination of ideas we have applied to other problems: it is possible to add or subtract numbers from all the lines which touch a particular point, without changing the shape of the minimum tour, and we can use the greedy algorithm to construct a minimum spanning tree for the changed distances. The minimum spanning tree is no longer than the shortest tour, since every tour is a spanning tree; but by properly modifying the distances we can make the minimum spanning tree very nearly a tour, so comparatively few possibilities need to be tried. I extended the Held and Karp method to take advantage of the fact that each point has at most four neighbors. In this way it was possible to verify at reasonable cost that this tour is optimum; but if I were faced with a larger problem, having say twice as many points to visit, there would be no known method to get the answer in a reasonable amount of time (*22*).

In fact, it may well be possible in a few years to prove that no good algorithm exists for the traveling salesman problem. Since so many people have tried for so many years to find a good algorithm, without success, the trend is now to look for a proof that success in this endeavor is impossible. It is analogous to the question of solving polynomial equations: quadratic equations were resolved in ancient Mesopotamia, and the solution of cubic and quartic equations was found at the beginning of the Renaissance, but nobody was able to solve arbitrary equations of the fifth degree. Finally, during the first part of the 19th century, Abel and Galois proved conclusively that there is no way to solve fifth degree equations in general, using ordinary arithmetic (*23*). It is now believed that there is no good algorithm for the general traveling salesman problem, and we are awaiting another Abel or Galois to prove it.

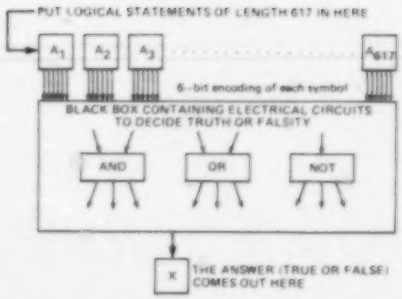


Fig. 11. Electrical network to decide the correctness of logical statements containing 617 characters or less.

In support of this belief, several important things have already been proved, notably that the traveling salesman problem is computationally equivalent to hundreds of other problems of general interest (*24*). If there is a good algorithm for any one of these problems, which for technical reasons are called NP-complete problems, then there will be good algorithms for all the NP-complete problems. Thus, for example, a good algorithm for the traveling salesman problem would lead immediately to a good algorithm for many other difficult problems, such as the optimum scheduling of high school classes, the most efficient way to pack things into boxes, or the best Steiner trees connecting a large number of points. A good solution to any one of these problems will solve them all, so if any one of them is hard they all must be.

A Provably Harder Problem

In recent years, certain problems have, in fact, been shown to be intrinsically hard, in the sense that there will never be a fast way to solve them. Probably the most interesting example of this type was developed in 1974 by A. Meyer and L. J. Stockmeyer (*25*). The problem is to decide whether or not certain logical statements about whole numbers 0,1,2,.... are true or false, even when the form of these statements is severely restricted.

Here are some examples of the sort of statements we must deal with.

$$048 \leq 1063$$

a statement which is clearly true.

$$\forall n\, \exists m(m < n+1)$$

This is logical shorthand that can be translated as follows, for people who are not familiar with the new math: "For all numbers n there exists a number m such that m is less than $n + 1$." It is clearly a true statement, since we may take m equal to n.

$$\forall n\, \exists m(m+1 < n)$$

"For all numbers n there exists a number m such that $m + 1$ is less than n." This statement is false, for if $n = 0$ there is no number less than zero; we are considering only statements about nonnegative numbers.

The next example is a little more complicated.

$$\forall x\, \forall y(y \geq x+2 \rightarrow \exists z(x < z \wedge z < y))$$

"For all numbers x and all numbers y, if y is greater than or equal to $x + 2$ then there exists a number z such that x is less than z and z is less than y." In other words, if y is at least 2 more than x, there is a number z between x and y, and this is obviously true.

Finally we can also make statements about sets of numbers; for example

$$\forall S(\exists x(x \in S) \rightarrow$$
$$\exists y(y \in S \wedge \forall z(z \in S \rightarrow y \leq z)))$$

"For all sets S of numbers, if there exists a number x such that x is in S then there exists a number y such that y is in S and for all numbers z in S we have $y \leq z$." Informally, the statement says that every set of numbers which is not empty has a smallest element, and this is true.

The logical statements we shall be concerned with cannot be essentially any harder than these examples; they may not involve subtraction, multiplication, or division; they cannot even involve addition except addition of a constant. (They cannot involve the formula $x + y$.) Thus the statements must be very simple—much, much simpler than those used every day by mathematicians constructing proofs of theorems.

According to a well-known theorem of Büchi (*26*), it is possible to decide in a finite number of steps whether or not any statement of the simple kind we have described is true or false, even though these logical statements may concern infinitely many cases.

But the new theorem says that it is impossible actually to do this in the real world, even if we limit ourselves to statements that can be written in no more than 617 symbols: "No realistic algorithm will ever be able to decide truth or falsity for arbitrary given statements of length 617 or less."

In order to understand exactly what this theorem means, we have to know what it means to speak of a "realistic" algorithm. The theorem of Meyer and Stockmeyer is based on the fact that anything that can be done by computer can be done by constructing an electrical

network, and so they envisage a setup like that shown in Fig. 11. At the top of such a device, one can insert any statement whose truth is to be tested. The logical language involved here makes use of 63 different symbols, including upper and lower case letters and a blank symbol, so we can place the statement (followed if necessary by blanks) into a sequence of 617 positions. Each position is converted into six electrical pulses, whose configuration of "on" and "off" identifies the corresponding character; thus, the letter X might be represented by the six pulses "off, on, on, off, off, on." The resulting 6×617 pulses now enter an electrical network or "black box" consisting of AND, OR, and NOT circuits; AND produces a signal that is "on" only when both inputs to AND are "on," OR produces a signal that is "on" when either or both of its inputs are "on," and NOT changes "on" to "off" and vice versa. At the bottom of the network, a pulse comes out which is "on" or "off" according to whether the given logical statement of length 617 was true or false.

According to Büchi's theorem, it is possible to construct such an electrical network with finitely many components, in a finite amount of time. But Meyer and Stockmeyer (25) have proved that every such network must use at least 10^{125} components, and we have seen that this is much larger than the number of protons and neutrons in the entire known universe.

Thus it is hopeless to find an efficient algorithm for this finite problem. We have to face the fact that it can never be done—no matter how clever we may become, or how much money and energy is invested in the project.

What should we do in the face of such limitations? Whenever something has been proved impossible, there is an aspect of the human spirit that encourages us to find some way to do it anyway. In this particular case, we might try the following sneaky approach: We could build an electric circuit which gives the correct answer in all simple cases and which gives a random answer, true or false, in the other cases. Since the problem is so hard, nobody will be able to know the difference.

But this is obviously unsatisfactory. A better approach would be to distinguish between levels of truth; for example, the answer might be "true," "false," or "maybe." And we could give various shades of "maybe," saying perhaps that the statement is true in lots of cases.

Let us consider the traveling salesman problem again. It is reasonably likely that, some day, somebody will prove that no good algorithm exists for this problem. If so, that will be a truly great theorem; but what should we do when we actually need to solve such a problem?

The answer, of course, is to settle for a tour that is not known to be the shortest possible one, but is pretty close. It has recently been observed that we can quickly find a traveling salesman's tour that is guaranteed to be no worse than 50 percent longer than the shortest possible tour, if the distances satisfy the triangle inequality. And algorithms have recently been developed for other problems that give answers which are probably correct, where the degree of probability can be specified, but the answer is not certain.

In this way, computer scientists and mathematicians have been learning how to cope with our finite limitations.

Summary

By presenting these examples, I have tried to illustrate four main points.

1) Finite numbers can be really enormous, and the known universe is very small. Therefore the distinction between finite and infinite is not as relevant as the distinction between realistic and unrealistic.

2) In many cases there are subtle ways to solve very large problems quickly, in spite of the fact that they appear at first to require examination of too many possibilities.

3) There are also cases where we can prove that a fairly natural problem is intrinsically hard, far beyond our conceivable capabilities.

4) It takes a good deal of skill to decide whether a given problem is in the easy or hard class; but even if a problem does turn out to be hard there are useful and interesting ways to change it into one that can be done satisfactorily.

References and Notes

1. G. Gamov, *One, Two, Three, . . . Infinity* (Viking, New York, 1947).
2. M. A. Morrison and J. Brillhart, *Math. Comput.* **29**, 183 (1975).
3. L. Euler, *Verh. Uitg. Zeeuwsch. Genoot. Wet. Vlissingen* **9**, 85 (1782); *Leonardi Euleri Opera Omnia* **7**, 291 (1923).
4. G. Tarry, *Mathésis* **20** (Suppl.), 23 (1900).
5. H. F. MacNeish, *Ann. Math.* **23**, 221 (1922).
6. M. Hall and D. E. Knuth, *Am. Math. Mon.* **72** (part 2, *Computers and Computing*), 21 (1965).
7. C. Tompkins, *Proc. Symp. Appl. Math.* **6**, 195 (1956); L. J. Paige and C. Tompkins, *ibid.* **10**, 71 (1960).
8. E. T. Parker, *ibid.* **15**, 73 (1963).
9. For a complete survey see J. Dénes and A. D. Keedwell, *Latin Squares and Their Applications* (Academic Press, New York, 1974).
10. See, for example, M. N. Barber and B. W. Ninham, *Random and Restricted Walks* (Gordon & Breach, New York, 1970), chap. 7.
11. W. A. Mozart, *Musikalisches Würfelspiel* (Schott, Mainz, 1956), K 516 f Anh. C 30.01; this was first published in 1793.
12. The 11 possibilities for bar 8 are all identical, and Mozart gave only two distinct possibilities for bar 16, so the total number of waltzes is 2×11^{14} rather than 11^{16}.
13. T. H. O'Beirne, *Dice-Composition Music* (Barr & Stroud, Glasgow, 1967).
14. E. W. Dijkstra, *Numer. Math.* **1**, 269 (1959).
15. See, for example, N. Christofides, *Graph Theory, An Algorithmic Approach* (Academic Press, London, 1975), sect. 7.4.
16. First construct the matrix of distances between all pairs of points, then try all possible intermediate junction points.
17. A determinant formula that specifies the number of spanning trees in a particular graph was discovered by C. W. Borchardt [*J. Reine Angew. Math.* **57**, 111 (1860)]. When the graph is a square grid, with m rows and n columns, the number of spanning trees seems to be always of the form mx^2 or $2mx^2$, where all the prime factors $> m$ of x are rather small numbers of the form $km \pm 1$; at least this is true when $m \leq 12$. For example, the large number cited in the text corresponds to the case $m = 11$, and in factored form it equals $2^{15} \cdot 11^2 \cdot 23^2 \cdot 89 \cdot 109 \cdot 199 \cdot 241 \cdot 373 \cdot 397 \cdot 419$. This curious circumstance, which I noticed while preparing the present article, still awaits a theoretical explanation.
18. J. B. Kruskal, Jr., *Proc. Am. Math. Soc.* **7**, 48 (1956).
19. E. W. Montroll, in *Applied Combinatorial Mathematics*, E. F. Beckenbach, Ed. (Wiley, New York, 1964), sect. 4.4.
20. This is the well-known "Hungarian method" for the assignment problem; for example, see (*15*, sect. 12.4).
21. M. Held and R. M. Karp, *Math. Prog.* **1**, 6 (1971).
22. It appears to be substantially easier to solve traveling salesmen problems when the distances are asymmetric, in the sense that the distance from x to y is uncorrelated with the distance from y to x. Such problems have been solved for up to 200 points by T. H. C. Smith, V. Srinivasan, and G. L. Thompson, *Ann. Discr. Math.*, in press.
23. See, for example, C. B. Boyer, *A History of Mathematics* (Wiley, New York, 1968), pp. 555 and 641.
24. R. M. Karp, in *Complexity of Computer Computations*, R. E. Miller and J. W. Thatcher, Eds. (Plenum, New York, 1972), p. 85. See also A. V. Aho, J. E. Hopcroft, J. D. Ullman, *The Design and Analysis of Computer Algorithms* (Addison-Wesley, Reading, Mass., 1974), chap. 10. For a popular account of related work, see G. B. Kolata, *Science* **186**, 520 (1974).
25. L. J. Stockmeyer, *The Complexity of Decision Problems in Automata Theory and Logic* (report MAC TR-133, Massachusetts Institute of Technology, Cambridge, 1974), chap. 6.
26. J. R. Büchi, *Z. Math. Logik Grundl. Math.* **6**, 66 (1960).

27. The preparation and publication of this article was supported in part by NSF grant DCR 72-03752 A02, ONR contract NR 044-402, and the IBM Corporation. Some of the computations were done with the MACSYMA system, supported by the Defense Advanced Research Projects Agency under ONR contract N00014-75-C-0661; others were done on the SUMEX-AIM computer, supported by NIH RR00785; still others were done at the Université de Montréal, Centre des Recherches Mathématiques, where a preliminary version of this article was prepared under the auspices of the Chaire Aisenstadt. A lecture based on this material was presented at the AAAS annual meeting in Boston, 22 February 1976, in the session entitled "The Frontiers of the Natural Sciences" organized by R. M. Sinclair.

200

Physical Limits in Semiconductor Electronics

Robert W. Keyes

This issue of *Science* attests to the revolution created by the advent of silicon microelectronics. The impact of semiconductor electronics in so many areas is a direct result of its record of providing ever-increasing information processing power per unit of cost. The steadily decreasing cost per digital operation has been achieved by fabricating more and more components—diodes, transistors, capacitors, and resistors—on a single piece, or chip, of silicon. A very primitive approximation asserts that the cost of producing a chip is independent of what devices or circuits are created on it. The number of components that can be fabricated on a single chip has increased from 1 to 30,000 over the last 15 years (*1*). Thus, it is not hard to understand that semiconductor electronics has not only made very large digital systems possible, but is also to be found in such cost-conscious markets as those for automobiles and home entertainment.

Progress to larger substrates, or chips, and miniaturization of components and interconnections have provided the keys to placing more components on a chip (*1*). Modern integrated electronics is based on planar technology, which means that devices and circuits are fabricated by operating on a surface of a chip: modifying it chemically by introducing impurities through masks and depositing layers of conducting and insulating substances in selected regions. The size of a chip is limited by defects, the existence of conditions that prevent the fabrication of an operable component in a certain region. A chip must be small enough to make the probability that it contains no defect reasonably large. Improvements in silicon substrates and in processing techniques have steadily reduced the density of defects per unit area and made progress to ever-larger chips possible (*2*). Chips measuring 0.5 by 0.5 centimeter are now common. Miniaturization involves decreasing all of the linear dimensions of structures fabricated on a chip, such as the widths of interconnection lines and the diameters of openings through which impurities are introduced by diffusion or ion implantation. One wonders how long this kind of progress can continue and whether and at what point laws of physics that limit further advances will be encountered. This article describes the search for such physical limits. Physics offers no reason to anticipate that increases in the size of chips through decreases in the number of defects per unit area will not continue. Advances in size are likely to come slowly, though, as major changes in the size of the substrate that must be processed may require major investments in new process tools.

Lithography

Continued miniaturization depends on improvements in lithography, the art of producing patterns in films that can be used as masks in processing—for example, to define patterns on a substrate where material may be altered by etching or by introduction of impurities. Lithography has received a substantial amount of attention as a limit to miniaturization (*3*). The most common form of lithography is photolithography, in which exposure to light changes the properties of a film of a photosensitive compound, allowing it to be selectively removed for purposes of masking. Exposure to light cannot, in practice, produce patterns with minimum dimensions much less than the wavelength of the exposing light. Thus, not too many years ago, dimensions of 1 or 0.5 micrometer were viewed as the limiting "least count" of microstructures.

Exposure of resists with electron beams has, however, emerged as a method that can produce much finer structures than are possible with optical exposure (*4*). In fact, there appears to be no significant fundamental limit to the resolution that can be achieved with electron beams; single atoms can be seen with the electron microscope (*5*). There are, however, practical limits. A very large number of resolution elements must be exposed to form a microstructure. For example, a chip 3 millimeters square is not large in the light of modern technology, but to expose it with 1-μm resolution implies exposure of 10^7 elements. A feasible production process must expose each of the 10^7 elements in a very short time. The highest possible beam current is desirable.

Thus, a presently accepted view of the limits to electron beam lithography runs as follows (*6*). Spherical aberration in electron lenses spreads the electron beam by an amount proportional to the cube of the incidence angle (α) accepted at the target (see Fig. 1). The diameter of the exposed spot is at least

$$D = C\alpha^3/2 \qquad (1)$$

where C is a constant that characterizes the spherical aberration of the electron lens. The current density in the beam is limited by the brightness of the source, B amperes per unit area per steradian. The current in the beam is proportional to the solid angle accepted, or to α^2, and thus large currents and small dimensions conflict according to Eq. 1. The electrons in the beam arrive randomly in time and are distributed in any specified time t according to Poisson's law. To ensure that statistical fluctuations do not leave any resolution element underexposed, it is necessary that the average exposure of an element exceed some minimum number of electrons, N. For example, if $N = 20$, then there is a probability of .006 that an exposure element receives only 10 electrons. Asking that the probability of an exposure less than half the average be smaller than 10^{-14} requires that N be something like 200. These limits on the beam current and on N imply that there is a minimum exposure time per spot, which depends on the spot diameter. The relation between the exposure time and D is shown in Fig. 1. As the cost of an electron beam exposure tool must be spread over a large number of components, Fig. 1 implies an economic limit on the applicability of electron beam lithography. Only very rough values of the parameters involved are needed to quantify this limit because of the very strong dependence on D. For example, taking the cost of the exposure system as $100 an hour and the value of the exposed silicon as $10 per square centimeter, an exposure rate $R = 3 \times 10^{-3}$ cm^2/sec is required, and it is found that $D > 5 \times 10^{-6}$ cm.

It would be a mistake to regard such a

The author is a research staff member at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.

201

limit on electron beam exposure technology as "fundamental," however. One possible escape might be found in the suggestion that it is possible to completely eliminate the effects of spherical aberration (7). Clearly, the invention of brighter electron sources could directly affect the exposure time. The potential of electron beam lithography is far from exhausted.

There is, though, another very important limit to the resolution of electron beam lithography as practiced today. Electrons pass through the resist, are scattered in the underlying silicon, and may emerge some distance away from the incident beam to expose the resist there (4). The range of a typical 25-kilovolt electron in silicon is about 3 μm, so there is a nonnegligible degradation of resolution by this backscattering effect. The effect can be diminished by reduction of the energy of the exposing electrons. Decreasing the electron energy, however, increases the chromatic aberration of the electron lenses, introducing another source of spreading of the spot.

The large-angle scattering that causes electrons to be returned from the substrate to the photoresist could be avoided by exposing with more massive particles—protons or other ions. The low brightness of ion sources has prevented the useful exposure of resists with ions to date. However, the development of more intense ion sources may be another path to pushing back the present limits on lithography.

X-ray lithography is an attempt to take advantage of the very high resolution that is possible in principle in writing with electron beams without being restrained by the long times that are needed to expose a chip by sequential scanning (3, 8). A high-resolution mask made by writing with an electron beam can be used to expose photoresist on a silicon substrate with x-rays; the short wavelength of the x-radiation, 1 to 100 angstroms, preserves the high resolution of the mask. Since a mask can be used to expose a large number of substrates, a long time and low current can be used to achieve high resolution in the preparation of the mask. The resolution attainable with x-rays is limited by a different effect: a secondary electron is emitted when an x-ray photon is absorbed, and the secondary electron has a range of one to several hundred angstroms, and exposes the resist in an area with this radius.

Thus, it must be emphasized again that the limits of particle beam lithography just described are in no way fundamental; they involve practical and economic considerations. Indeed, fabrication of

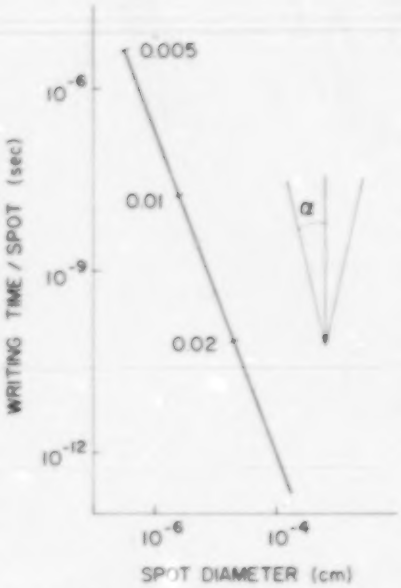


Fig. 1. Minimum time needed to expose a picture element of an electron resist with a focused electron beam as a function of element diameter (6). The limit is calculated with the following parameters, which are typical of modern electron beam systems: spherical aberration, $C = 5$ cm; source brightness, $B = 10^6$ amp/cm^2-steradian; minimum number of electrons per spot, $N = 200$. The values of α, the beam angle accepted in Eq. 1, are indicated on the limiting line.

metal conductors 80 Å wide has been reported (9). One can be confident that the march of lithography to smaller dimensions will continue. Physical limits to miniaturization must be sought elsewhere.

Hot Electrons and Breakdown

Another problem of miniaturization of digital devices arises from the fact that voltages cannot always be reduced in proportion to dimensions, so that electric fields in semiconductor devices increase with decreasing size. Voltage levels in digital systems must be large enough to provide a clear distinction between "on" and "off" states of a device, or between informational 0's and 1's. Digital signals must be standardized; a 0 must look like a 0 and a 1 must look like a 1, regardless of their source.

The uncertainties caused by processing defects and randomness in the distribution of impurities also must be overcome by the application of large voltages. Like the thermal voltage, these requirements are insensitive to the size of devices (10). In its application to electrical signals, this means, for example, that if it is intended that a voltage V_1 represents a 1, then a circuit that receives a voltage V_1' less than V_1 must transmit a voltage that is closer to V_1 than V_1' is. Otherwise, there would be a steady deterioration of voltage level as information passes from stage to stage through the processor. When it is realized that a similar statement must apply to voltages near V_0 that are intended to represent 0, it is clear that the standardization requires a nonlinear response.

Voltages applied to electrical devices change the potential of electrons in some spatial regions with respect to those in other regions. Now, electrons are dispersed in energy by an amount of approximately κT by thermal agitation (κ is the Boltzman.. constant and T is absolute temperature). Electrical voltages that are small relative to $\kappa T/q$ are thus just a small perturbation on the steady energy distribution of the electrons and produce only linear effects (q is the charge of an electron; $\kappa T/q = 0.025$ volt at $T = 300$ K). Nonlinear effects can be produced by voltages that are large relative to $\kappa T/q$.

This scale of nonlinearity is most perfectly exemplified by the ideal p-n junction, in which the current depends on the voltage as

$$i \propto \exp(qV/\kappa T) - 1 \qquad (2)$$

The junction characteristic as expressed by Eq. 2 is a practical limit to electical nonlinearity. A similar scale even seems to be applicable in biology: neuron voltages are a few times $\kappa T/q$. Various lines of thought agree in establishing $\kappa T/q$ as a practical minimum voltage scale for the production of nonlinear electrical effects. The word *scale* implies that the attainment of the very large nonlinearities needed for reliable logical operation in the presence of such disturbing influences as cross talk, environmental fluctuations, and component variability will require that voltages a great many times $\kappa T/q$ be used in real circuits. Questions such as the necessary degree of reliability and the acceptable amount of component variability lie outside the realm of quantitative physical science, and the only statement that can be made about the voltage is that it must be large relative to $\kappa T/q$.

Thus, the historical trend for voltage levels in semiconductor circuits to decrease with the progress of miniaturization cannot long continue, and voltages must become relatively independent of size. Nevertheless, all dimensions of a structure, including such internal device dimensions as base widths in bipolar transistors and the thicknesses of the depletion layers between p- and n-type regions of silicon, are decreased as miniaturization advances. Electric fields and high-field phenomena such as hot

electrons and dielectric breakdown grow in importance with decreasing dimensions and form a limit to miniaturization.

A quantitative version of one such limit has been formulated (*11*). Consider the field-effect transistor (FET) shown in Fig. 2. The depletion regions associated with the *p-n* junctions of the source and drain electrodes are shown. The source-drain distance must be greater than the sum of the widths of the depletion layers in order that the gate can exercise control of the conductance along the surface. The width of the depletion layers can be reduced by increasing the doping level of the substrate silicon. The source and drain can then be placed closer together and the transistor made smaller. However, the electric fields in the depletion layers will be increased at any given applied voltage, and the voltage limit of the transistor will be decreased. The same is true of the depletion layer that is formed between the surface and the bulk substrate silicon when a conductive surface charge is induced by a voltage applied to the gate. In this case, the increase in doping results in an increase in electric field in the oxide. Thus, both junction breakdown and oxide breakdown limit the miniaturization of FET's. Since breakdown of silicon and of SiO_2 is a well-studied subject, it is possible to calculate for each voltage the maximum doping level that can be used and the smallest permissible source-drain separation. A calculation of this type is shown in Fig. 3 (*11*). The results presented show that breakdown in the oxide insulator is the limiting factor in the miniaturization of FET's. However, curvature of a junction and nonuniform doping profiles affect the breakdown voltage but are difficult to take into account quantitatively, and it has also been suggested that breakdown at the drain-substrate junction is the more serious limitation.

Related models can be developed for bipolar transistors. The basic requirement is that the base shall not be completely depleted or suffer "punch-through," and that the junctions shall not break down. The punch-through is controlled by heavy doping, which, however, decreases the breakdown voltage of the junctions and the voltage at which the transistor can operate.

The electron temperature, a rough measure of the average electron energy, rises above the lattice or thermal equilibrium temperature at electric fields much smaller than those needed to cause avalanche breakdown. The electrons become "hot." Some of the hot electrons have enough energy to pass from the silicon into the SiO_2 insulator on the surface and become trapped there, simulating a potential applied to the gate and changing the properties of the surface (*12*). Such effects are noticeable in both field-effect and bipolar transistors (*12*, *13*). In FET's hot electrons produced by electric fields in the channel at the surface change the threshold voltage when they escape into the insulating SiO_2. Hot electrons that are produced in the region close to the intersection of a *p-n* junction with the silicon surface increase the leakage current across the junction along the surface when they escape into the SiO_2 and degrade the performance of bipolar transistors. Quantitative studies of such effects in FET's are available and show that undesirable changes in characteristics occur rapidly if the fields in the channel are only slightly greater than 10^4 volt/cm (*12*). Although much more experimental information is needed, that which is available suggests that these hot electron effects will limit reduction of the FET "length," the source-drain separation, to something like ¼ μm. Information concerning the hot electron degradation of *p-n* junctions is even more scanty.

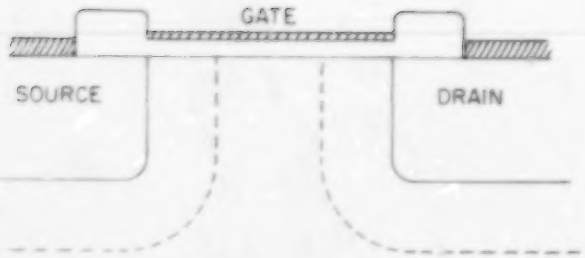


Fig. 2. Structure of an insulated gate field-effect transistor, showing the extent of the depletion regions associated with the source and drain junctions.

Power Dissipation

An issue closely connected to miniaturization of devices is density. Density means the number of components or number of circuits per unit area. High density is desirable for several reasons. The time taken for a signal to propagate from one circuit to another is reduced as the density is increased. Generally speaking, too, the cost of producing a structure is reduced as the area that it occupies decreases. The most serious limitation on the density of logical processing circuits, if the necessary lithographic techniques have been mastered, is set by power dissipation.

Thus, we turn to another set of problems that are associated with the processing of digital information. Although it seems possible to construct nondissipative information processing systems in principle (*14*), in practice known kinds of information processing devices dissipate power, converting it to heat. There are very basic reasons for the dissipation of power in logical processors. Most important is the irreversible nature of information processing in a general-purpose computer. Information is represented in physical degrees of freedom, such as the charge on capacitors. In information processing different starting points may lead to the same result; the memory of the initial conditions is lost, and there is no way to reverse the physical processes employed to return to the initial state of the system. The logical irreversibility implies physical irreversibility and dissipation (*15*). Although in principle one can preserve a complete history of the computer operations, so that they could all be reversed and all of the energy used eventually recovered (*14*), this is not practical in a modern electrical general-purpose computer. In such a computer, logical processing stages perform their function, pass the information on to the next stage, and are restored to a state in which they are ready to process information again, losing their memory of what was done in the preceding step. Thus, each logical circuit converts a certain amount of electrical power to heat. The heat must eventually be transferred to some fluid, commonly air or water, that carries it out of the system. The heat usually leaves the semiconductor chip itself by conduction across a surface to another solid and is transferred to a fluid across some larger

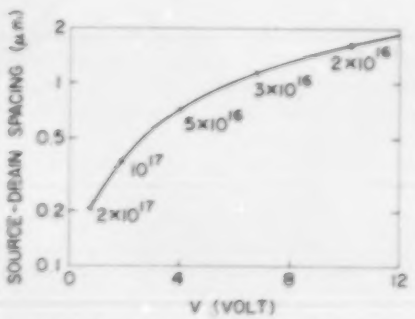


Fig. 3. An example of the limits imposed on a metal oxide semiconductor field-effect transistor as shown in Fig. 2 by breakdown in the SiO_2 insulator (*11*). Heavy doping reduces the widths of the depletion layers illustrated in Fig. 2; the impurity concentrations needed to achieve particular source-drain spacings are indicated. Breakdown in the oxide then limits the supply voltage of a simple inverter to the value shown.

Table 1. Parameters achieved in packaging an air-cooled computer processing unit with 100,000 circuits (*17*).

Parameter	Value
Average power per circuit	65 mwatt
Circuit delay on chip	0.6 nsec
Dissipation density	0.4 watt/cm^2
Circuits per chip (average)	57
Chip density	0.12 cm^{-2}
Circuit density	70 cm^{-2}
Signal velocity	10^{10} cm/sec

area. There is, however, a finite limit to the rate at which heat can be transferred across any surface, and therefore, a limit to the density at which components that dissipate power can be packed on a planar surface.

At one level the problem of power dissipation is intimately bound to miniaturization. Energy is dissipated in the successive charging and discharging of capacitances that are part of semiconductor devices. Capacitance is dimensionally dielectric constant times a linear dimension and so must scale in proportion to dimension. The energy in the capacitor is proportional to the square of the voltage, V, applied to it. If a circuit element is operated at a rate $1/t$, then the power dissipated in it has the form

$$P = AV^2L/t \qquad (3)$$

where A is a proportionality constant with the dimensions of a dielectric constant and L is a length parameter, say the square root of the area per element. Then limitation of the dissipation per unit area to a value Q which can be removed by heat transfer means that

$$P/L^2 < Q \qquad (4)$$

Equations 3 and 4 immediately lead to limiting relations between S and t and between P and t. I believe that circuit considerations, the need to send logic signals from one component to another efficiently over transmission lines, limit A to something like 1000 times the dielectric constant of free space (*16*). The limits that result from this assumption are shown in Fig. 4.

This thermal limit can be qualitatively understood in the following way. The energy supplied to the circuit is used to charge capacitances. Capacitance is dimensionally electric permittivity times a linear dimension. As circuits are miniaturized, the energy per operation therefore can be reduced in proportion to linear dimension. The number of circuits per unit area, however, increases inversely as the square of linear dimension, and the power per unit area thus increases inversely to dimension at a constant operating rate. Eventually the power density becomes greater than the maximum rate of removal Q; this is the density represented by the solid line in Fig. 4.

A limit also arises from the finite time taken for a signal to propagate from one circuit to another. The higher the circuit power, the farther apart the circuits must be placed to provide area for the transfer of heat, and the longer the propagation time. An estimate of the propagation time limit is also given in Fig. 4.

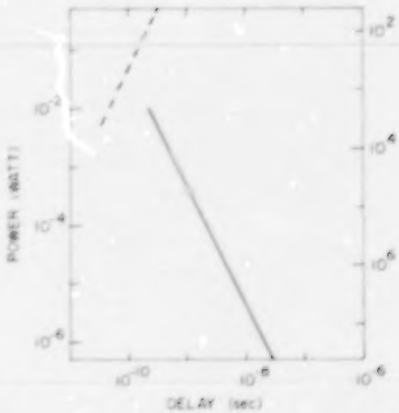


Fig. 4. The thermal limit on logic circuits following from Eqs. 3 and 4, $Pt^2 > (AV^2)^2/Q$ (solid line), calculated for $A = 10^{-10}$ farad/cm, $V = 1$ volt, $Q = 20$ watt/cm^2 (*16*). The propagation time limit, $Pt^{-1} > c_s^2 Q/m^2$ (dashed line), is also shown for the parameters: velocity of transmission, $c_s = 5 \times 10^9$ cm/sec; length of transmission divided by square root of area per circuit, $m = 10$. The number of circuits per square centimeter corresponding to the power is shown on the right.

Packaging

Actually, propagation times become more important at a higher level of packaging. The preceding discussion, in which the circuits are regarded as densely packed on a substrate, is applicable to semiconductor chips. The chips are mounted on larger carriers, frequently made of ceramics, when they are assembled into a larger system. The chip carriers and the substrates or boards on which they are mounted provide for mechanical and electrical attachment of the chip and contain the wiring matrix that interconnects the chip and the power distribution conductors (*17*). Simple mechanics may limit the closeness with which chips may be placed on a large substrate, through problems such as making the chip carrier conveniently replaceable for fault correction or design changes. On the other hand, heat dissipation can also limit the density of chips. The rate at which heat can be removed from boards is much less than the rate at which it can be removed from chips. Chips are cooled by conduction through an interface to some larger solid, and the heat current is limited by the thermal resistance of the solids and the interface; values of 20 watt/cm^2, as used in Fig. 4, can be attained. At the board level, however, heat is transferred to air, which carries it out of the system; 1 watt/cm^2 is difficult to achieve over a large area. Thus the chips may have to be widely spaced to provide for cooling, and the chip-to-chip propagation time can become an important limit on performance.

The delay encountered by a signal that must pass from one chip to another is the sum of two parts, the on-chip circuit delay and the chip-to-chip propagation delay. As just described, if the circuits are driven at a higher speed more power is used, and the chips must be placed farther apart, increasing the propagation delays. The art of mounting the chips on a substrate in such a way that they can be interconnected and cooled is called packaging. Packaging problems comprise one of the most severe limitations on the performance of modern high-speed integrated logic. The many functions that must be taken into account in the design of a package are:

1) Provide mechanical support and attachment.

2) Provide electrical connection to chip.

3) Transform chip contact dimensions to mechanically pluggable dimensions.

4) Contain wiring matrix for chip interconnections.

5) Contain power distribution network.

6) Receive heat from chip and deliver it to a fluid.

7) Protect the semiconductor from the environment.

The emergence of fast devices has required that interconnection lengths be reduced to take advantage of the device speeds. Packages must be made smaller. Thus problems of heat dissipation and mechanical access to the chip to replace failed units or to incorporate design changes have become much more severe. Some of the parameters that have been achieved in a recent large computer are presented in Table 1 (*17*).

Memory

The preceding discussion has been implicitly oriented toward logic circuits.

The same considerations apply, however, to memory, with certain differences in detail. Like logic, memory is digital, and reading, writing, and moving information requires the application of voltages large enough to produce nonlinear effects and is a source of the power dissipation that accompanies the large voltages. Magnetic memories, cores and bubbles, are "nonvolatile" and can preserve information for a long time without dissipating any power. To a degree, semiconductor memories that store information in the charge on a capacitor share the same characteristic, although the stored charge slowly leaks away and must be restored periodically. Active use of these memories, though, involves reading the information through some kind of matrix addressing, or transferring it from site to site until it arrives at some place where it may be read and transmitted to other parts of a system. These nonlinear operations produce heat and subject memory devices to the thermal limitations described above. The principal differences between memory and logic are that memory elements often enjoy a low duty cycle; that is, they are infrequently actively involved in an operation. They are also much simpler physically than logic gates, and so a memory cell occupies less area on a silicon chip than a logic circuit. The result of the lower dissipation of memory cells and their smaller areas is that their dissipation per unit area is not much different from that of logic circuits. Memory is subject to essentially the same thermal limitations as logic, as illustrated in Fig. 4.

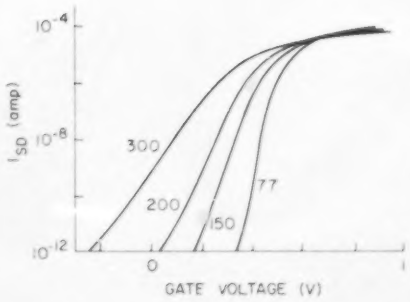


Fig. 5. Source-drain current of a silicon field-effect transistor as a function of gate voltage and temperature, illustrating the sharpening of characteristics as the temperature is lowered (*18*). The curves are labeled with the temperature in Kelvins.

Alternatives

The problems and limitations of silicon electronics have led to a search for new directions that might relax the limits described. Several possibilities have received attention. One of these is abandoning semiconductor electronics in favor of superconducting electronics. A second involves operating silicon devices at low temperatures, typically 77 K, the boiling temperature of liquid nitrogen (*18*). Physics immediately suggests many possible advantages fo operating silicon devices at low temperatures. Most obvious is the increase in conductivity; the scattering of electrons by lattice vibrations decreases as the temperature is lowered. The effect of the decreased resistance is probably most important in the metallic interconnections of integrated circuits, since conductivity in the silicon itself is primarily limited by scattering by impurities rather than lattice vibrations. Low temperatures also promise to reduce the power dissipated. As explained above, the sharp transitions between states of a device that are the essence of digital electronics require that voltages large compared to the thermal voltage, $\kappa T/q$, must be used. Thus, the lower the temperature, the lower the voltage required in switching circuits. Figure 5 shows dramatically how the sharpness with which an FET is turned on by the gate potential increases as temperature is decreased. The power dissipation is expected to be proportional to the square of the voltage (*18*). Thus, these rough considerations suggest that 16 times less power will be required at 77 K than at 300 K.

Low temperature may lead to even greater decreases in power in memories in which information is stored in charge on a capacitor. The charge is retained by opening a transistor switch; it gradually leaks away as reverse current through some *p-n* junction and must be periodically refreshed. The reverse current of junctions decreases rapidly with decreasing temperature, so refreshing is needed less frequently at low temperatures. In some circuits the power dissipated in the refreshing operations is an appreciable part of the total dissipation and may be reduced by a large amount (*18*).

The decreased power dissipation means that less area is required to allow for removal of heat. The lowered resistance of metals allows connecting lines to be made narrower and also decreases areal requirements. Thus, significantly higher device densities promise to be achievable at low temperatures. The cost of low-temperature semiconductor electronics should be substantially smaller than that of conventional semiconductor devices. The unanswered economic question is, How large a system is required to justify the added investment in refrigeration by reduction in cost and increase in performance of semiconductor devices?

Another new direction in electronics is the development of digital integrated circuits in other semiconductors. In fact, this direction is not really new, as hope that semiconducting compounds of group III and group V elements would replace silicon and germanium in transistor applications has been entertained ever since the discovery of the compounds more than 20 years ago (*19*). A large share of the enthusiasm for the III-V semiconductors was based on their high electron mobilities, which are in many cases 2 to 20 or more times greater than mobilities in silicon. The search for physical limits reveals one reason why the III-V semiconductors have failed to displace silicon; the limits have little to do with mobility, but concern such things as avalanche breakdown fields and thermal problems. It is necessary to have an energy gap that is large enough to prevent intrinsic conductivity, the thermal excitation of electrons from the valence band to the conduction band, from interfering with device operation. The favorable features of a semiconductor are a high energy gap, which allows the temperature of a component to rise a certain amount and decreases the stringency of the cooling requirements, and high breakdown fields, which are associated with a large gap. A semiconductor with an energy gap smaller than that of silicon is unlikely to play an important role in modern electronics.

An even more important reason for the dominance of silicon, however, was the rapid establishment of a feasible processing technology for silicon. The ease with which an SiO_2 layer can be formed on a silicon surface and the remarkable properties of such a layer as an insulator, a diffusion mask, and a neutralizer of undesirable surface effects are unmatched by any phenomena in the III-V compounds.

Nevertheless, the III-V compounds have had an impact on electronics. The advantages of gallium arsenide as a material for transistors have long been known. It has a larger energy gap than silicon, slightly larger breakdown fields, and a much higher electron mobility. The more difficult technology of GaAs has at last been mastered, and it has become the superior material for microwave transistors (*20*). Field-effect transistors are used to take advantage of the high electron mobility; in bipolar transistors both electon and hole mobilities are important. One other semiconductor, indium phosphide, appears to share the advan-

tages of GaAs, but its technology is still in a more primitive stage.

The technology that makes GaAs microwave transistors possible is now being extended to the fabrication of integrated microwave circuits. One wonders when the application of the proved high-speed microwave capabilities of GaAs to digital circuits will begin. Exploratory attempts have already been made (*21*). A realistic view suggests, however, that the invasion of large-scale digital applications by GaAs will be much more difficult than the conquest of the microwave field. The reason is that one or a few microwave devices with superior frequency response can extend the bandwidth of a system and have great economic value. On the other hand, digital systems use thousands to millions of devices, and low-cost fabrication is essential. Fast devices are less important because the speed of a system is also limited by the delays in the package. The highly developed and versatile silicon technology, optimized by many years of experience, will not be displaced easily.

Summary

Although the limitations of the methods of lithography in use at a particular time are easily recognized and attract substantial attention, experience shows that technological ingenuity keeps pushing them to ever-smaller dimensions. There seems to be no fundamental reason to expect that lithographic limits will not continue to recede. The limits to the advance of miniaturization are to be found in the ability of materials to withstand high electric fields and in the ability of packaging technology to remove heat from active components and provide for power distribution, signal interconnection, and flexible mechanical assembly.

References and Notes

1. G. E. Moore, in *Technical Digest 1975 International Electron Devices Meeting* (IEEE, New York, 1975), pp. 11–13.
2. J. L. Buie, in *1976 WESCON Technical Papers* (Western Electric Show and Convention, El Segundo, Calif., 1976) paper 23/2.
3. G. Pircher, in *Solid State Devices, 1975* (Société Française de Physique, Paris, 1975), pp. 31–72; J. T. Wallmark, in *Solid State Devices, 1974* (Institute of Physics, London, 1975), pp. 133–167.
4. G. R. Brewer, *IEEE Spectrum* **8** (No. 11), 23 (1971).
5. A. V. Crewe, J. Wall, J. Langmore, *Science* **168**, 1338 (1970).
6. T. E. Everhart, in preparation.
7. N. W. Parker, S. D. Golladay, A. V. Crewe, in *Scanning Electron Microscopy/1976*, O. Johari, Ed. (IIT Research Institute, Chicago, 1976), part 1, pp. 37–44.
8. H. I. Smith and S. E. Bernacki, *J. Vac. Sci. Technol.* **12**, 1321 (1975); R. Feder, E. Spiller, J. Topalian, *ibid.*, p. 1332.
9. A. N. Broers, W. W. Molyen, J. J. Cuomo, N. D. Wittels, *Appl. Phys. Lett.* **29**, 596 (1976).
10. R. W. Keyes, *Applied Physics* **8**, 251–259 (1975); *IEEE Trans. on Computers*, in press.
11. B. Hoeneisen and C. A. Mead, *Solid State Electron.* **15**, 819 (1972); *ibid.*, p. 981.
12. S. A. Abbas and R. C. Dockerty, *Appl. Phys. Lett.* **27**, 147 (1975).
13. D. R. Collins, *ibid.* **13**, 264 (1968).
14. C. H. Bennett, *IBM J. Res. Dev.* **17**, 525 (1973).
15. R. Landauer, *ibid.* **5**, 183 (1961); *Ber. Bunsenges. Phys. Chem.* **80**, 1048 (1976).
16. R. W. Keyes, *Proc. IEEE* **63**, 740 (1975); *IEEE J. Solid State Circuits* **SC-10**, 181 (1975).
17. For a detailed description of the packaging of a modern computer see R. J. Beall, in *1974 INTERCON Technical Papers* (IEEE, New York, 1974), paper 18/3.
18. F. H. Gaensslen, V. L. Rideout, E. J. Walker, in *Technical Digest 1975 International Electron Devices Meeting* (IEEE, New York, 1975), pp. 43–46; F. H. Gaensslen, V. L. Rideout, E. J. Walker, J. J. Walker, *IEEE Trans. Electron Devices*, in press.
19. D. A. Jenny, *Proc. IRE* **46**, 959 (1958).
20. C. A. Liechti, *IEEE Trans. Microwave Theory Tech.* **MTT-24**, 279 (1976).
21. R. L. Van Tuyl and C. A. Liechti, in *1974 IEEE International Solid State Circuits Conference Digest of Technical Papers* (Winner, New York, 1974), pp. 114–115.

Solid-State Electronics: Scientific Basis for Future Advances

J. A. Giordmaine

The leaps forward in conceptual understanding, the new device principles, the advances in analytical technique, and the achievements in materials preparation that make up the scientific basis for the electronics revolution described in this book can be readily identified, in retrospect. A list of such contributions would certainly include crystal structure analysis based on x-ray and electron diffraction, the explanation of conductivity in terms of the quantum theory of solids, the growth of ultrapure single crystals of electronic materials with controlled doping, the concept of a semiconductor amplifier, the invention of high-frequency oscillators based on stimulated emission, and the demonstration of quantum tunneling devices.

As Niels Bohr remarked, however, it is very difficult to predict, especially the future. Any attempt to identify a scientific basis for future advances is limited by a number of constraints. The lead time between scientific discovery and utilization in solid-state technology is at least 5 years, frequently more than enough time for the economic factors determining utility to have changed beyond recognition. The important scientific advances rarely emerge in a completely scheduled or planned pattern and not always in response to a perceived need. Progress often occurs in stages of abrupt change in the conceptual basis of the field, following a steady if undramatic accumulation of essential background understanding. Today the challenges of large-scale integration and the pace of change of the technology are making heavy demands on resources previously dedicated to longer-term research. Fundamental research of all kinds is carried on in a climate of increasingly critical scrutiny and diminishing real support.

In spite of the maturing of semiconductor technology and the uncertainties and vulnerability of the research enterprise, I believe that there is considerable ground for optimism about the prospects for continuing innovation in solid-state electronics. From a fundamental point of view, our present degree of control over electrons and their motion in solids may be compared with our ability to manipulate light at the beginning of the 19th century. We are now able to exploit behavior dependent on electron density and current flow, analogous to quantity of optical radiation and radiant intensity. The device utilization of the wave nature of electrons—making specific use of amplitude, phase, and coherence as in the case of light in diffraction, interference, holography, and the laser—has barely begun with the discovery of the Josephson effect.

From a nearer-term point of view, the exponential growth of the scale of integration provides new openings for the invention of functional substitutions—ways of using properties of matter to

The author is director of the Solid State Electronics Research Laboratory, Bell Laboratories, Murray Hill, New Jersey 07974.

achieve results otherwise obtained by brute force multiplication of devices. Examples of such functional substitutions are optoisolators, quartz-crystal filters, and optical fiber interconnection of computers—economical solutions that drastically reduce the number of components required to perform a given function. Closely related examples are magnetic bubbles and charge-coupled devices, which combine multiple-device functions in a single monolithic structure. The spread of integrated circuit applications has led to a proliferation of specialized "custom" circuits, on a scale much larger than the variety of earlier discrete transistors. More economical, mass produced, general-purpose circuits, programmable by electrical or other means after manufacture, may become increasingly important. Such circuits and their processing or programming represent a merging of hardware and software functions in circuit architecture that will make new types of demands on semiconductor technology.

A substantial increase in the speed of logic and memory access into the subnanosecond region would have a fundamental impact on digital data and signal processing in computers and communication. Extrapolations based on scaling arguments, however, suggest that current silicon technology can yield only modest potential increases in device speed. Moreover, basic limits on heat removal by liquid cooling rule out the utilization of silicon technology in high-density circuits at much higher speeds than are presently available. The need for increases in digital processing speed offers a natural occasion for technological innovation. The recently introduced small-scale optical integrated circuits containing injection lasers coupled to modulators and other components by optical waveguides, as well as medium-scale superconducting integrated circuits based on Josephson tunneling devices, are examples of radically new directions for which the scientific basis is undoubtedly still emerging.

In this article I will cite a few examples of areas of research that can reasonably be expected to contribute to future advances in solid-state electronics.

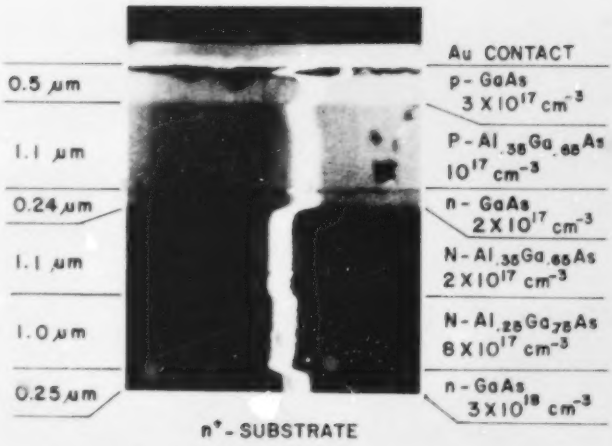


Fig. 1. Scanning electron micrograph of an MBE double-heterostructure laser. The active *n* region has a thickness of 0.24 μm; the superimposed line scan is a secondary emission signal to help delineate the layers. [Photograph courtesy of A. Y. Cho, R. W. Dixon, H. C. Casey, Jr., and R. L. Hartman]

Surfaces, Thin Films, and Epitaxy

The early contributions of surface science to solid-state electronics are perhaps epitomized by the essential role played by the understanding of surface states in the work on metal-semiconductor interfaces that immediately preceded the transistor (*1*). The theory (*2*) of the influence of surface states on rectification characteristics provided the working basis needed to deal with the unexpected results of preliminary experiments on field-effect amplification, and in fact opened the way to the discovery of point contact transistor action in 1947. As part of the expansion of surface physics work during the subsequent two decades, a detailed understanding of the electronic properties of a few well-characterized surfaces in high vacuum, including silicon, emerged from the measurements of band bending, work function, and photoemission spectra (*3*). Metal oxide semiconductor (MOS) technology, which arose during the mid-1960's and early 1970's, depended in a fundamental way on the control of silicon–silicon oxide interface phenomena, including surface states, surface charges, and space charges in the oxide. It should be noted that the success of this technology to date, which required detailed characterization of the interfaces and solution of difficult instability problems due in part to ion mobility in the oxide, was achieved in the absence of a basic understanding of the atomic arrangement or the electronic structure of the interface.

Prospects for providing such a fundamental picture have increased in recent years as a powerful combination of improved experimental and theoretical tools has been brought to bear on the study of clean surfaces. On the experimental side, combinations of different electron spectroscopies (*4*) applied to clean, well-characterized surfaces in ultrahigh vacuum have allowed accurate determination of surface state energy levels both near and within the band gap. For example, observations of the same silicon surface by ultraviolet or x-ray photoelectron spectroscopy, two-electron ion-neutralization spectroscopy, electron loss spectroscopy, Auger electron spectroscopy, and low-energy electron diffraction provide independent information on states both in the vicinity of the surface and directly at the surface, on the chemical composition of the surface, and on the symmetry of the surface lattice arrangement.

On the theoretical side, new computational methods are available for calculating detailed maps of the electron distribution at a surface, taking the atomic arrangement into explicit account self-consistently (*5*). The surface geometry is either a relaxed version of the bulk structure or a reconstructed arrangement with symmetry wholly different from the bulk. A comparison of the electron spectroscopy results with theoretical predictions of surface state energy spectra based on a particular surface model allows confirmation of the assumed atomic arrangement. Together, the theoretical and experimental approaches have provided the first complete description of surfaces, including identification of the atomic species present, their atomic arrangement, and the distribution of valence electrons in space and energy. The recent extension of the calculations to GaAs-Ge junctions (*6*) suggests that these methods will be applicable to semiconductor and metal-semiconductor interfaces as well, and should yield an atomically detailed picture of the interface structure, electron states, charges, reconstruction, and the related junction electronic properties. An arsenal of recently introduced experimental tools, including angle-resolved photoelectron spectroscopy, synchrotron far-ultraviolet and x-ray spectroscopies, and backscattering and channeling of energetic ions, suggests a substantial rate of progress.

A beginning has also been made in the investigation of polycrystalline semiconductor thin films (*7*). An understanding of the behavior of grain boundaries in such films, for example, is not now avail-

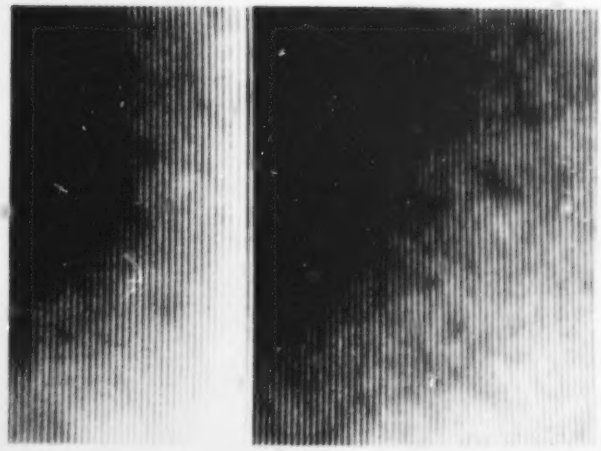


Fig. 2. Bright-field transmission electron micrograph of MBE alternating layers of GaAs and AlAs; the periodicity of the pattern is 56 Å; at the right of the interface the structure is seven layers of GaAs and three layers of AlAs; at the left of the interface the structure is nine layers of GaAs and one of AlAs. [Photograph courtesy of P. M. Petroff]

able, and will be essential in exploring their potential for low-cost solar cells.

A particularly promising example of an outgrowth of basic surface science research is molecular beam epitaxy (MBE) formation of compound semiconductor films (*8*). In this technique separate atomic and molecular beams from multiple thermal sources in high vacuum irradiate a substrate at intensities chosen to grow films with a desired composition and doping. The slow growth rate together with the independent control of the separate beam sources allows fabrication of semiconductor junction profiles, both in doping and in composition, to a precision approaching the level of a single atomic layer. The MBE process originated in fundamental studies of the adhesion of gallium and arsenic atoms to GaAs surfaces; early characterization of MBE surfaces was aided by the unique opportunity for electron diffraction diagnostic probing of the surface in situ during film growth in vacuum. Molecular beam epitaxy has already been used to prepare films and layer structures for a variety of GaAs and $Ga_xAl_{1-x}As$ devices. These include varactor diodes having highly controlled hyperabrupt capacitance-voltage characteristics, IMPATT diodes, microwave mixer diodes, Schottky barrier field-effect transistors (FET's), injection lasers (Fig. 1), optical waveguides, and integrated optical structures. The potential of MBE for future solid-state electronics is greatest for microwave and optical solid-state devices and circuits where submicrometer layer structures are essential, and where the inherent adaptability of the process to planar technology and integration will offer a number of opportunities. A measure of the potential of the technique for millimeter wave electronics is a recently demonstrated MBE GaAs Schottky barrier diode cryogenic mixer with noise temperature of 315 K at 102 gigahertz (*9*).

Possible longer-term implications of MBE for solid-state electronics are related to the capability of growing extended layer sequences with alternating composition such as GaAs and AlAs (Fig. 2). Such superlattice structures with periodicities of 50 to 100 angstroms show negative resistance characteristics attributed to resonant tunneling into the quantized energy states associated with the narrow potential wells formed by the layers (*10*). Detailed optical spectra and electron microscope studies of MBE superlattices establish that the potential well distributions can be controlled and positioned to a precision of a few atomic layers.

Electron diffraction and optical studies have been made of MBE multilayers prepared by exposing a GaAs substrate sequentially to beams of gallium, arsenic, and aluminum in brief pulses carefully timed to deposit the equivalent of a single atomic layer (*11*). For a narrow range of substrate temperatures around 850 K, a monolayer crystalline compound forms with regular periodicity extending over domains of at least 100 Å in size, having the planar structure GaAsAlAsGaAs . . . and an equivalent formula $AlGaAs_2$. The new material differs from the parent crystals GaAs and AlAs and from the solid solution in its optical and electronic properties as well as its short-range and long-range order. The stability of such synthetic new materials, grown at temperatures considerably below the melting point of the solid solution, may, in fact, be determined by kinetic rather than thermodynamic considerations. The technique may make available a new class of electronic materials similar in composition to naturally occurring compounds but distinct in structure and properties.

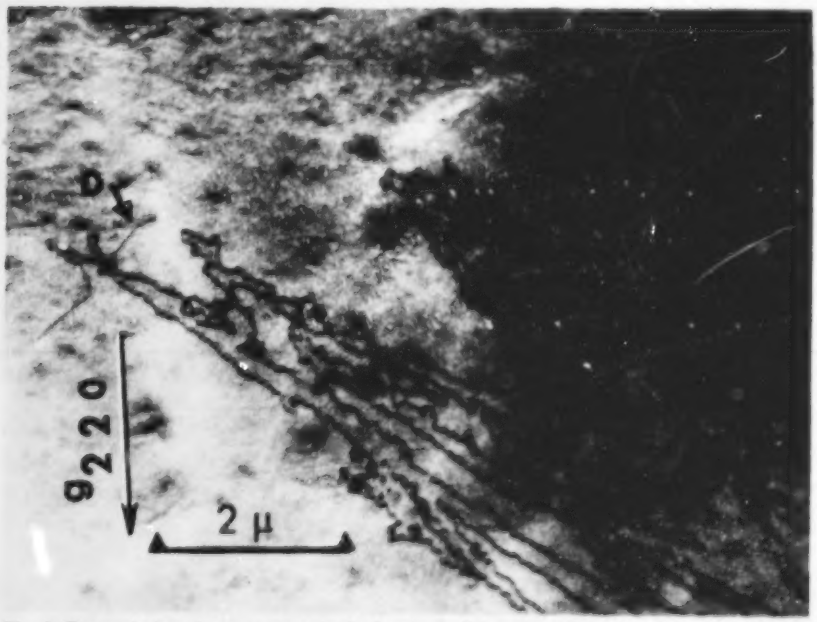


Fig. 3. Transmission electron micrograph showing a dislocation (*D*) at the origin of a dark line defect network (g_{220} is a lattice direction). [Photograph courtesy of P. M. Petroff and R. L. Hartman]

Defects, Recombination, and Reliability

As silicon technology has reached its present state of maturity, the reliability of integrated circuits has become largely determined by residual defects distributed randomly over the chip area. The number of components per chip has

meanwhile been increasing by a factor of about 2 each year. The result of the reduction in device size has so far been an improvement in reliability per individual component or per circuit function. With further miniaturization expected from submicrometer electron beam and x-ray lithography, however, this picture can be expected to change. Practical limits on size set by reliability problems, as well as the limits on size, packing, and circuit configuration set by power dissipation, will be felt long before fundamental limits set by thermodynamic or quantum considerations become important (*12*). The reliability limits arise in large part from the mobility of atoms, vacancies, dislocations, and other defects driven by electric fields and gradients of temperature or chemical potential. The problems include electromigration, electrochemical corrosion, and surface and bulk diffusion of metallization; they already appear in present devices and integrated circuits under special conditions. There is still very little fundamental understanding of these processes. A broader picture of the physics of solid-state diffusion will be essential to take full advantage of future advances in lithography.

In technologies other than silicon, the reliability constraints are quite different. In the case (discussed below) of Josephson junction circuits operating at liquid helium temperature, for example, not only have heat dissipation problems disappeared, but there are as yet no known mechanisms for circuit degradation. That favorable situation is balanced, however, by reliability problems in thermal cycling due to severe mismatch in thermal expansion of the substrates and metallization films.

An example of a technology in which defect motion already appears to have important consequences for reliability is optoelectronics. In GaAs and other light emitting diodes (LED's) and injection lasers made with compounds of group III and group V elements, light is emitted by radiative recombination of electrons and holes in the vicinity of a junction or heterojunction. Whenever a nonradiative recombination event takes place, however, the available band-gap energy is released to the lattice in a burst equivalent to an instantaneous local temperature of $\sim 10^4$ K. When this process occurs at a vacancy or interstitial defect, the available energy is often enough to activate motion of the defect (*13*). The mobile defects are trapped at any nearby dislocation and can cause the dislocation to grow (*14*). In injection lasers and LED's such recombination-induced motion may give rise to growth of the dense, opaque networks of dislocations termed dark line defects which ultimately cause premature failure (*15*) (Fig. 3). Improved lifetimes of present communication lasers and LED's are the result of careful growth conditions, control of defect concentration, and avoidance of dislocations on which mobile point defects can collect (*16*). A more complete characterization of defects, their motion, and their interactions with dislocations will provide a basis for future, more permanent solid-state optical devices.

A new spectroscopic technique, deep level transient spectroscopy (DLTS) (*17*), allows direct measurement of energy levels and concentration of the electrically active defects in the neighborhood of a junction. DLTS observations also measure the cross section of the defect for stimulation of nonradiative recombination. Such results have led to useful models of the coupling of the defect to the lattice, and new information on how the energy released in a nonradiative process is transferred to phonons and finally appears as heat in the crystal (*18*). Only a few of the defects characterized by DLTS have so far been identified. The identification of larger numbers of such defects will be essential for full control of luminescent devices.

Defects also play a well-known but poorly understood role in the parasitic high-frequency oscillations sometimes observed in aged injection lasers, in the stability and noise figure of microwave FET amplifiers, and in other solid-state electronics applications. Present studies of defects are providing an understanding of a variety of mechanisms of device degradation. The future manufacture of more reliable optoelectronic and microwave devices will rest on such an understanding.

Superconductivity and Speed

Early attempts to utilize the phenomenon of superconductivity for solid-state logic and memory devices made use of its most striking property: the abrupt loss of electrical resistance at the phase transition from normal to super-

Fig. 4. Probe measurement of phase-dependent normal and pair potentials along a superconducting strip near a constriction. Scale bar, 1 μm. [Photograph courtesy of G. J. Dolan and L. D. Jackal]

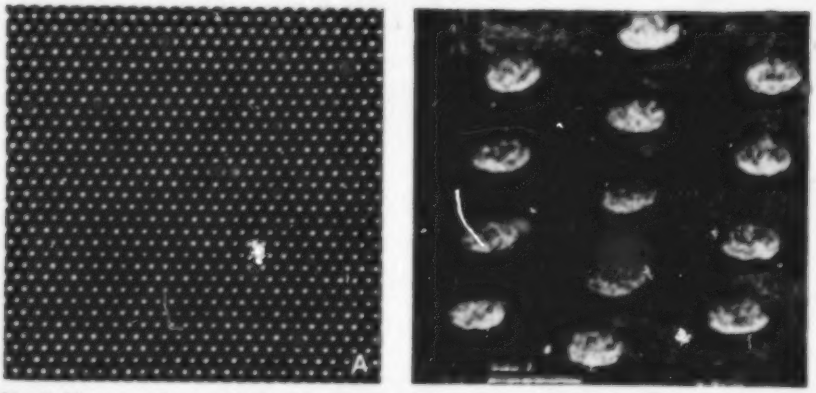


Fig. 5. Electron micrograph (B) showing a two-dimensional triangular lattice pattern of 1-μm holes in superconducting aluminum film. When the magnetic flux density is adjusted to match the flux vortex lattice with the hole lattice, strong "resonant" vortex pinning is observed. (A) Structure in (B) shown at lower magnification. [Photograph courtesy of A. F. Hebard, A. T. Fiory, and S. Somekh]

conductivity. The best known of these devices, the cryotron (*19*), consists of a metallic gate film kept superconducting at a temperature slightly below its transition temperature, and an adjoining control film. The gate film can be converted to its normal state by a magnetic field due to current in the control film. The two stable conduction states of the gate film and the switching action of the control film provide memory and logic functions. The switching time of this device was expected to be about 10^{-9} second, determined by eddy current damping in the normal phase during propagation of the phase boundary down through the submicrometer film thickness. In practice, however, the switching time turns out to be 10^{-5} to 10^{-7} second, determined either by a lateral propagation time due to nonuniform phase nucleation or by a thermal time constant associated with transfer of a residual latent heat of transition. This device is no longer competitive in performance with room-temperature devices and the cryotron has not had widespread application.

The conceptual leap forward that has led to a fresh look at the possibility of a superconducting electronics technology originated in fundamental solid-state physics research, rather than a direct attempt to make superconducting devices. This idea was, of course, the recognition that superconducting electron pairs, as well as normal electrons (*20*), could tunnel from one superconductor to another across a thin-film insulator (Fig. 4).

The predicted behavior, soon discovered in the laboratory (*21*), gives rise to two states of conduction across an insulating tunnel junction, one a normal dissipative current driven by finite voltage across the junction, the other an electron pair or Josephson supercurrent occurring at zero voltage. Switching between the two voltage states by magnetic field or current injection from adjoining control lines (*22*, *23*) provides the basis for memory and logic. Unlike the cryotron, the Josephson junction is switched entirely in the superconducting phase and avoids delays of phase boundary propagation. The inherent delay of the Josephson switching action is determined primarily by the superconducting energy gap, a value corresponding to a time of $\sim 10^{-12}$ second (1 psec). Indirect experimental measurements show a delay time of ≤ 5 psec. The high speed (10- to 100-psec logic delay compared to $\sim$ 1 nanosecond for high-speed transistors) and low switching power (10 microwatts compared to 10 to 100 milliwatts for high-speed transistors) make Josephson technology an attractive candidate for high-performance computing and data processing.

One of the main problems requiring further attention before this technology can find any wide application is the thermal cycling damage. While Josephson junctions may well have an indefinite operating life at liquid helium temperature, cycling of the structures to room temperature introduces severe stress as a result of the different thermal expansion of the metallization layers, the insulating oxide layer, and the substrate. Stress relaxation tends to occur by hillock growth on the films, short-circuiting the 40- to 50-Å-thick oxide insulating layer on repeated cycling. Fundamental research on the metallurgy, oxidation, tunneling, deposition, and growth of superconducting materials could determine the future of this promising technology (Fig. 5).

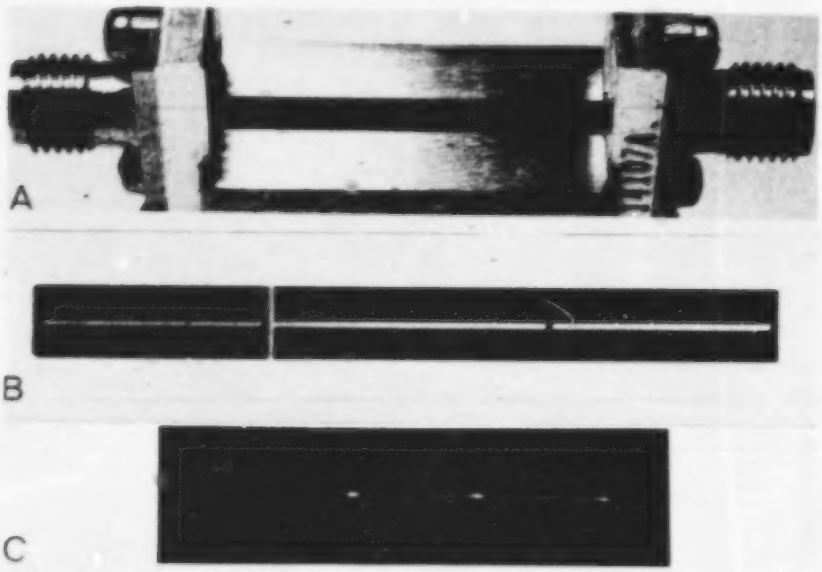


Fig. 6. (A) Picosecond-speed solid-state switch formed from a 1-inch metallic stripline on a 10-mil silicon substrate. The right terminal of the strip is charged. A 10-psec optical pulse (2×10^{-9} joule) illuminates the 8-mil gap at the right to generate a pulse moving toward the left; the crossbar shorting stubs act to form a rectangular pulse. The gap at the left is irradiated by a second optical pulse for correlation measurement of the electrical pulse length. (B) Enlargement of stripline, reversed. (C) Two-photon fluorescence photograph of a picosecond pulse train. [Photographs (A) and (B) courtesy of D. H. Auston, C. Shank, and E. Ippen]

Materials and Quantum Electronics

Research on materials, crystal growth, and solid-state optics is becoming an increasingly visible segment of quantum

electronics. Research on solid-state laser sources, largely confined in the past to GaAs and the GaAs-Al,Ga,As heterostructure system, has been extended to a wide range of binary, ternary, and quaternary III-V and IV-VI compounds (*24*). The potential importance of the 1- to 1.1-μm region has stimulated work on junction lasers based on ternary-ternary, ternary-quaternary, and other heterostructures such as Ga,As,Sb-Al,Ga,As,Sb; InP-Ga,In,As,P; and In,GaAs-In,Ga,P. The binary and ternary IV-VI compounds have become increasingly important as sources for the 3- to 30-μm region (*25*). The Pb,Sn (Te,Se) systems are unusual in that the energy gap passes through zero within the ternary composition range and makes possible a continuous range of infrared laser sources.

The complex materials requirements for a double-heterostructure injection laser arise from the need for both carrier and optical confinement, to avoid dislocation-prone lattice mismatch at the heterostructure interface, and for low levels of nonradiative centers. Although a large body of information on phase diagrams, dependence of band gap on composition, and certain other properties of these complex systems is becoming available (*24*), an enormous amount of further data and understanding will be needed before reproducible, low-cost production of reliable devices becomes feasible, especially outside the GaAs-Al,Ga,As system.

The materials science effort in this field has made possible improvements in and new combinations of liquid phase epitaxy and MBE growth, and in turn a wide range of new injection laser techniques. Recent advances include distributed feedback lasers, with a periodic index variation adjoining the active region to provide sharp frequency selectivity; Bragg reflector lasers, with the usual cleaved end mirrors replaced by coupled passive optical waveguides containing etched Bragg reflection sections; and many different types of stripe lasers. Virtually all continuously operating lasers are now constructed in a narrow stripe geometry to provide sharper transverse mode selection and better coupling of the laser light output to an optical fiber.

As the position of optics in communications and solid-state electronics becomes more established, there will be a developing need for combinations of optical functions, as in multiple sources and modulators, multiplexed optical sources, and perhaps optical sources with nonlinear devices to generate new or tunable frequencies. It is likely that there will be advantages of function, stability, and cost in fabricating these combinations on a single substrate, as is the case for silicon integrated circuits. In progress toward such integrated optical structures in recent years (*26*), surface waveguides have been demonstrated in many materials, on many substrates, and by many different techniques, including sputtering, evaporation, ion implantation, and diffusion. Lenses, couplers, and gratings, as well as active devices such as injection lasers and waveguide modulators, have been adapted to couple to optical waveguides and fit the integrated optics format. Simple small-scale integrated optical circuits have been demonstrated using both monolithic and hybrid techniques. Monolithic circuits, so far fabricated only in the GaAs-Al,Ga,As system, have included laser, waveguide, modulator, coupler, and Bragg grating elements. Hybrid circuits have been demonstrated with a wide variety of magneto-, acousto-, and electro-optical techniques and materials. A substantial amount of materials science and device research is being directed toward improved growth techniques and exploration of the possible original functions such circuits may perform.

Future advances in solid-state electronics can also be expected on the basis of the scientific and technical developments in ultrashort light pulse generation and nonlinear optics over the past 15 years. As an example, consider the recent use of ultrashort light pulses, generated by mode-locked solid-state and dye lasers, as the basis of a picosecond solid-state switch (*27*). Ultrashort visible light pulses, with energy in the nanojoule range, produce substantial surface conductivity at a silicon surface as a result of the creation of high-density plasma. The picosecond time scale of the onset of surface conductivity has been applied to ultrahigh-speed electrical switching and gating circuits. In the example of Fig. 6, the silicon photoconductor forms the dielectric substrate of a microstrip transmission line. A gap in the line can be closed at 10-psec rate by a visible light pulse; the line can be shorted by a subsequent infrared pulse, producing bulk photoconductivity. The use of standard transmission line reflection techniques allows generation of a sharply terminated pulse. Research in this field is closely coupled with materials studies of electron-hole diffusion and recombination under high-density plasma conditions.

Nonlinear optical techniques, such as optical harmonic generation, optical parametric oscillation, and stimulated Raman scattering, have so far had an impact primarily on research instrumentation. Widespread application of nonlinear optics in solid-state electronics—for example, its routine use in integrated optical circuits for frequency conversion—will probably require specialized materials with strongly nonlinear behavior. While there now appears to be a good working knowledge of the nonlinear properties of inorganic crystals, organic nonlinear susceptibilities are still only partially understood (*28*). Research on the mechanisms of the nonlinear optical properties of organic materials may well lead to materials with nonlinear optical properties superior to those of presently known inorganic crystals.

The examples cited above are perhaps a few of the plausible research directions likely to have an impact on the *evolution* of solid-state electronics during the coming years. A comparison of the variety and richness of the physics and materials science at hand today with that, say, of 1945 would seem to argue for at least as favorable prospects for an electronics *revolution* during the coming 32 years.

References

1. W. Shockley, *Electrons and Holes in Semiconductors* (Van Nostrand, New York, 1950), pp. 31–34.
2. J. Bardeen, *Phys. Rev.* **71**, 717 (1947).
3. F. G. Allen and G. W. Gobeli, *ibid.* **127**, 1500 (1962).
4. J. E. Rowe and H. Ibach, *Phys. Rev. Lett.* **32**, 421 (1974); J. E. Rowe, *Phys. Lett. A* **46**, 400 (1974); H. Hagstrum, *Science* **178**, 275 (1972).
5. J. A. Appelbaum and D. R. Hamann, *Phys. Rev. B* **6**, 2166 (1972); *Phys. Rev. Lett.* **31**, 106 (1973); *ibid.* **32**, 225 (1974); K. C. Pandey and J. C. Phillips, *ibid.*, p. 1433.
6. G. A. Baraff, J. A. Appelbaum, D. R. Hamann, *Phys. Rev. Lett.* **38**, 237 (1977).
7. K. J. Bachmann, E. Buehler, J. L. Shay, S. Wagner, *Appl. Phys. Lett.* **29**, 121 (1976).
8. A. Y. Cho and J. R. Arthur, *Prog. Solid State Chem.* **10**, 157 (1975).
9. A. Y. Cho, R. A. Linke, M. V. Schneider, in preparation.
10. L. L. Chang, L. Esaki, R. Tsu, *Appl. Phys. Lett.* **24**, 593 (1974).
11. A. C. Gossard, P. M. Petroff, W. Weigmann, R. Dingle, A. Savage, *ibid.* **29**, 323 (1975).
12. R. W. Keyes, *Proc. IEEE* **63**, 740 (1975).
13. D. V. Lang and L. C. Kimerling, *Phys. Rev. Lett.* **33**, 489 (1974).
14. L. C. Kimerling and P. M. Petroff, *Appl. Phys. Lett.* **29**, 461 (1976).
15. R. L. Hartman and P. M. Petroff, *ibid.* **23**, 469 (1973).
16. B. C. DeLoach, *Bell Lab. Rec.* **53**, 203 (1975).
17. D. V. Lang, *J. Appl. Phys.* **45**, 3023 (1974).
18. ——— and C. H. Henry, *Phys. Rev. Lett.* **35**, 1525 (1975).
19. D. A. Buck, *Proc. IRE* **44**, 482 (1956).
20. B. D. Josephson, *Phys. Lett.* **1**, 251 (1962); I. Giaever, *Phys. Rev. Lett.* **5**, 147 (1960); *ibid.*, p. 464.
21. P. W. Anderson and J. M. Rowell, *Phys. Rev. Lett.* **10**, 230 (1963); J. M. Rowell, *ibid.* **11**, 200 (1963).
22. J. Matisoo, *Proc. IEEE* **55**, 172 (1967).
23. T. A. Fulton, J. H. Magerlein, L. N. Dunkleberger, in "Proceedings of the 1976 Applied Superconductivity Conference," *IEEE Trans. Magn.* **MAG-13**, 56 (1977).
24. M. B. Panish, *Proc. IEEE* **64**, 1512 (1976).
25. T. C. Harman and I. Melngailis, *Appl. Solid State Sci.* **4**, 1 (1974).
26. P. K. Tien, *Rev. Mod. Phys.*, in press.
27. D. H. Auston, *Appl. Phys. Lett.* **26**, 101 (1975).
28. B. F. Levine, in *Dielectric and Related Molecular Processes*, M. Davies, Ed. (Chemical Society, London, in press), vol. 3.

Index

A

AEROSAT, 51
ALGOL, 16, 136
APD (*see* Avalanching photodiode detector)
APL, 125
APT, 16
ARPA computer network, 6
ATM (*see* Automatic teller machines)
ATS-6, 53
Academic Computing Group (ACCOMP), 134, 135
Advanced Research Projects Agency (ARPA), 136
Air Defense Computer (AN/FSQ-7), 82
Air defense system, computer needs, 82
Air-quality computations, 58–59
Airport surveillance radar (ASR-7), 83
American Express, "Space Bank" computer systems, 33
American Telephone and Telegraph Co., 96, 97, 114
  communications satellites, 51
  computer industry, 6, 15
  data terminals tariffs, 102
  digital data service, 82
  long lines, 98
  telephone industry leadership, 101
  Teletype Corp., 98
Angiography, 42
Antenna design, beam widths, and patterns, 52–53
Atlanta, Ga., fiberguide system experiment, 171
Atmospheric chemistry, modeling processes, 58
Atmospheric contaminants, computerized computations, 59
Automatic teller machines, terminals, 31
Automation
  acceptance in industry, 79–80
  advanced systems models, 80
  control systems economics, 80
  dehumanization effects, 119
  energy conservation, 79
  environmental, control processes, 79
  industry, 8
  psychology laboratory tests, 140
Automobile emission standards, 87
Automotive electronics
  alcohol ignition interlocks, 89
  controlled dual displacement engine (DDE), 86
  displays, 88
  electromagnetic interference, 88
  electronic fuel injection, 86
  engine knock sensors, 87
  history and development, 86
  low-cost components, 88
  microprocessors, 86
    large-scale integration technology, 86
    metal oxide semiconductors, 86
  multiplex systems, 88
  radar-augmented braking, 89
  solid-state ignition systems, 86
  subsystems, 87
  trucks, 87
Avalanching photodiode detector (APD), 168, 170

B

BASIC, 16, 36
BankAmericard Service Corp., computer credit card verification system, 32
Banks
  batch systems, financial use, 31, 32
  computers, 6
  electronic transfers, 6, 31, 32
Batch systems (*see* Banks)
Beam
  addressing, 183
  deflection dynamics, 183, 187
  storage, 183
Bell system, 101
Bell Telephone Laboratories, 56, 165
  electron beam exposure system (EBES), 187–188
  integrated circuits, 98
  library network, 57
  MERCURY system, 57
  optical communications, 171
  Pierce loop, 100
  telecommunications cable fabrication, 169
  UNIX time-share system, 60
BELLREL system, 57
Binary coded decimal, 138, 140
Biophysical Sciences Laboratory, 45
Bipolar logic, integrated injection logic, 181
blood pressure telemetry, 38
  See also Medical devices
British Broadcasting Corp., 6
Bubble memories
  development, 183
  fabrication, 182
  files, 182
  lattice file, 183
  operation, 182
  technology, 182
Bureau of Labor Statistics, Dept. of Labor, 110, 112–113

C

C3D (*see* Cascade charge-coupled device)
CAI (*see* Computer-aided instruction)
CAT (*see* Computerized axial tomography)
CCD (*see* Charge-controlled device)
COBOL, 16, 124
COMSAT (Communications Satellite Corporation), 47, 50
CPU (*see* Central processing unit)
CRT (*see* Cathode-ray tube)
CTD (*see* charge-transfer device)
Canadian communication satellites, 50, 51
Cardiovascular research, implantable telemetry systems, 38
  See also Medical devices
Cascade charge-coupled device, 37
Cathode-ray tube, 78, 81, 93
Central processing unit, 32, 33
  components, 139, 140, 144
  Intel 8080, 140, 145
  mathematics calculations, 142
  memory and storage, 140
  microcoded IMP-16, 143
  packaging, 200
  telephone directories, 163, 164
Charge-controlled device, 26
  access time comparisons, 182
  cost effectiveness, 182
  signal processing, 182
  storage density, 181
Charge-coupled devices, 181
Charge-transfer device
  applications, 173
  characteristics, 173
  delay lines, 172
  research, 172
  signal processing costs, 178
  structure, 173
  transversal filters, 176
Circuit components costs, 21
Circuits
  design-cost relationships, 22
  dimensions, 21
  large-scale integration, 4, 12, 18, 26, 81, 83, 93, 136, 144
  noniterated, 21
  photolithography, 20, 181
  semiconductor memories, 179
  small-scale integration, 26
  technology, 86
Citibank, electronic funds transfer system, 32
  See also Banks
Cleveland Clinic, 43
Communication Satellite Act, 1962, 47
Communication system, PLATO, 73
Communication Workers of America, 114
Communications
  digital channel transmission, 98
  federal regulations, 97
  fiber transparency, 9
  lasers, defect control, 205
  networks, switching, 78, 98
  science and technology, 57, 93
  See also Bell Telephone Laboratories
Communications satellites
  ATS-6, 53
  American Telephone and Telegraph Co., 51
  Anik, 50
  Canadian, 50, 51
  component reliability, 52
  COMSTAR, 50–51
  Courier B, 47
  data transmissions, 55
  Early Bird, 47, 48
  earth stations, 48–49, 53–54
  Echo I, 47
  electronics application, 51–52
  experimental systems, 53
  General Telephone and Electronics, 51
  global communications, 47, 49
  INTELSAT systems, 48–49
  ocean service, 51
  Project Relay, 47
  Project Syncom, 47
  Project Telstar, 47
  RCA, 50
  SATCOM, 50–51
  Space Shuttle, 53
  system trends, 53–54
  traveling wave tube amplifiers, 48, 51–52
  WESTAR, 50–51
  Western Union, 50
Communist Manifesto, 119
Computer-aided instruction, 132

Computer applications, 140–144
automatic teller machines, 31–32
batch systems, 31
consumer products, 33
control, 17
coronary care, 36
data processing, 16, 17
distribution and marketing of goods, 33
insurance, 33
minicomputers, 14, 15, 138
on-line inquiry terminals, 31
reservation services, 33
scientific calculations, 37
securities markets, 33
ultrasonic imaging systems, 37–38
x-ray tomography, 36
See also PLATO
Computer-based information system, 122
Computer circuits
commercial applications, 5
electron beams, 21
optical techniques, 20–21
resolution problems, 20–21
Computer communications
bandwidth availability, 99
Boolean logic, 98
cost effectiveness, 91, 96–97
digital networks, 100
IBM monopolies, 97, 102
modulation technology, 99
multiplexing, 99
new services, 96
packet switching networks, 99
Computer components
amplifiers, 27
assembly, 19
costs, 27
hardware, 122
integrated circuits, 19
manufacture costs, 19
memories, 27
microprocessors, 27
transducers, 27
Computer control theory, history and development, 79
Computer-controlled devices, 17
Computer-controlled process, 17
Computer-controlled systems, 76–77, 78
Computer devices
electric fields, 19
modems, 99
Computer education, PLATO system, 68–71, 72–73
Computer industry
American Telephone and Telegraph Co., 6, 15
antitrust cases, 106
assembly line techniques, 113
Beckman Instruments, 10
Bell Telephone Laboratories, 6, 23, 26, 56, 165
Burroughs, 17
calculator manufacturing, 110
data processing, 157
economics, 27, 28
effects on work, 118–119
France, 108
Germany, 108
growth limitations, 111
Hewlett-Packard, 10, 26, 63–65
human-machine interaction, 161, 165
IBM, 6, 16, 17, 25, 26, 96, 97, 98, 102
international competition, 105–108
Japanese exports, 105
labor demands, 109
marketing, 26
minicomputers, 14, 15, 138
Motorola, 25, 26
organization control, 119
privacy, 119
tectronics, 10
Texas Instruments, 25, 26
U.S. exports-imports, 106
Varian, 10
wafer manufacture, 26–27
Computer-laboratory equipment integration, 58
Computer networks, 13, 14
distributive, 129, 132, 135
equality of access, 14
line-switching, 136
manpower, 11
message switching, 136
organization, 133
Computer programs
ALGOL, 16
algorithms, 16
APL, 125
APT, 16
BASIC, 16
COBOL, 16
FORTRAN, 16
PL/1, 16
software, 16
Computer revolution, dignity of man, 120
Computer scanners, improvements, 46
Computer sciences
academic role, 28
engineering schools, 28
finite problems, 197
innovations, 56
research processes, 60
university research, 28
Computer systems, microprocessors, 22
Computerized axial tomography, 41, 42
scanners and scanning procedures, 42, 43, 44
Computerized x-ray scanners
cerebral images, 45
composition, 44
heart images, 45
need for speed, 44
neuroradiology, 43, 44
radiation doses, 42, 46
reconstructed images, 42, 43, 44–45
university research, 44, 45, 46
Computers
algorithms, 60
amino acid analyzer, 143
analog-to-digital converters, 58
applications, 140–144
artificial intelligence, 120
atmospheric measurements, 58
atmospheric ozone computations, 59
automatic operation, 12
banking industry, 6, 31, 32
biological monitoring systems, 141
business fraud potential, 120
calculations, 8
capabilities, 116
centralized management, 14
commercial marketing, 14, 15, 17
communications, 90, 91, 96–99, 100, 102
component technology development, 19, 27, 139
continuous process control, 16, 17
control systems costs, 78–80
core memories, 82, 179
cost effectiveness, 4, 19
costs, 14, 15
credit card transactions, 32
data processing, 16, 17
dehumanizing management, 151
digital, 11, 12, 98, 146
digital-to-analog converters, 58
distributed processing, 6
economics, 118
electronic funds transfers, 6, 31, 32
electronic memories, 179
ENIAC, 123, 125
federal government use, 4, 10
financial firms, 31, 32
heuristic procedures, 60
history, 15, 16
hobby shops, 14, 15
household appliances, 14
human behavior simulation, 120
IBM 650, 56, 57
IBM 704, 58, 82
IBM 705, 82
IBM 7094, 58
IBM OS/360, 156
industrial processes, 78
information
handling, 7, 16
management, 6, 121, 125–127, 151, 152, 154, 155
retrieval, 57, 90, 91, 153
sharing, 14
system processing, 122
inquiry and verification systems, 31, 32
instruction aids, 132
intelligent terminals, 14
interconnections, 18
languages, 136
large-capacity magnetic recording devices, 179
large finite numbers, 189–190
laser exposure printing, 93
Latin squares, 191
library retrieval tool, 56
management information systems, 17, 122, 151, 155
manpower, 4, 7
marketing, 31
mathematics, 54, 142, 144, 190–191, 192, 195
mean time between failures, 181
mechanized work, 118
medical administration, 5
memory
capacity, 181
technology, 123, 182, 183, 201
mental processes theories, 59
meteorological measurement, 59
microprocessor, 4, 5, 13, 14, 15, 22, 122, 138, 139, 149, 150
military support, 82–83
minicomputers, 4, 14, 15, 58, 138, 149, 150
modeling, 56, 58
multiprocessing systems, 13, 143
multiprogramming systems, 13, 16, 158–160
multivariate statistical analyses, 59
networks, 98, 194, 196
numbers calculations, 58
on-line service, 31, 32, 57
operating systems, 158
ownership distribution, 4
parallel processors, 13
perception research, 59
perfused biological samples experiments, 143
phototypesetter, 57, 93, 94
point-of-sales electronic registers, 31, 33, 34
polynomial equations, 195
Postal Service, 6
problem-solving, 191, 196
process control, 14
processing distribution, 98
processors, 143
program capability, 12
programming languages, 16, 123
psychological research, 59, 120
psychophysiology research, 59
publishing economics, 57
random paths, 192, 193
record keeping, 117
research tool, 56
research tools, 6

reservation systems, 31
robots, 13, 14
scientific calculations, 16
self-regulatory capability, 12
semiconductor memories, 179
semiconductors, 168, 172, 187, 197, 200–203
single-stream processors, 13
software development, 155, 156, 158, 160
solid-state circuits, 4, 9, 54, 202, 203, 206–207
speech
research, 47
synthesizers, 57, 58
statistical experiments, 191
storage
capability, 12, 184–185
devices, 123, 179, 183
technology advancements, 122
urban atmospheric studies, 59
U.S.-U.S.S.R. competition, 157–158
WHIRLWIND system, 81
Computers-communications interaction, 6, 8, 57, 97, 98
COMSAT General Corp.
MARISAT system, 51, 54, 55
satellite communication system, 51
Core memories, 89
core technology, 179
credit cards computerized transactions, 32
history and development, 179
production, 179
Cybernetics, 17
Cyclotron, radionuclide images, 46

D

Data base management systems (DBMS), 17, 122
logical storage management, 124, 125
physical storage management, 124, 125
Data communications systems, 11
Data information, satellite transmission, 55
Data processing, 16, 97, 146
applications, 17
data base management systems, 17, 122
digital data, 85
education-experience practices, 157
history and development, 157
international competition, 157
management information systems, 17, 122, 151, 155
quality-quantity problems, 157
U.S.-Japan competition, 105–107, 158
Deep level transient spectroscopy (DLTS), 205
Department of Commerce, electronics trade figures, 105
Department of Transportation, Canada, 51
Digital devices miniaturization
density problems, 199
dielectric breakdowns, 199
electric fields, 198, 199
power dissipation, 199, 200, 201
Digital transmission, glass fibers, 168, 170

E

EAM (*see* Electric accounting machine)
EFTS (*see* Electronic funds transfer systems)
EMI (*see* Electromagnetic interference)
EPA (*see* Environmental Protection Agency)
Early warning/tracking systems, 82
Education
electronics media, 28, 61
instructional technologies, 61, 62, 64, 122, 132
on-line displays, 74, 76
PLATO, 68–71, 72
technology applications, 63–65
Educational communications (EDUCOM), 133
Electric accounting machine, computer transition, 17
Electrical devices, voltage levels, 198
Electromagnetic interference scanner, 43
Electron beam
accessed memories, 183
exposure system operation, 188
high-resolution photomasks, 21
lithography, 21, 186, 187, 197, 198
Electronic control systems, cost effectiveness, 91
Electronic devices
acoustics, 5
calculators, 3
calculators-educational impact, 5
diagnostic devices, 42
INTELSAT amplifiers, 52
magnetic recording devices, 179
medical equipment, 5, 35, 36, 37–39, 41, 42
memories, 123, 179, 183, 184
microprocessors, 13, 14, 15, 122, 123, 146, 148
military uses, 5, 23
solid-state, 4, 9, 202, 203, 206–207
transistors, 4, 23
Electronic fuel injection, 86
Electronic funds transfer systems, 31, 32, 90–91
Electronic industry
Japanese technology advancements, 105, 107
semiconductor memories, 180
United States
domestic consumption, 105
economics, 110
military sales, 107
Postal Service employment, 113
productivity decrease, 110
Electronic mail system
character encoding, 92
characteristics, 91
economics, 91, 94
Mailgrams, 90
need, 91, 94
optical scanning terminal, 92
privacy protection, 92
teleprinter systems, 90
telex/TWX, 90
volume, 92
Electronic printing systems, 93, 94
Electronics
American manufacturers, 26
automotive, 86
balance of trade, 105
circuitry costs, 78
communication satellites, 51–52
controlled dual displacement engine (DDE), 86
definitions, 8
digital devices miniaturization, 198, 199
digital processors, 93
educational innovations, 61
electron beam, 21, 93, 183, 186–188, 197, 198
electron spectroscopy comparisons, 203
federal government
regulations, 10
support, 10
human factors engineering, 162
industrial competition, 105, 106, 107
industrial innovations, 10
invasive instruments, 36
large-scale integration, 4, 12, 18, 81, 83, 93, 136, 144, 181
logic circuits, thermal limit, 200
medical use, 35
memories, 179, 183, 201
military needs, 81, 83
miniaturization, 186
noninvasive instruments, 37, 41
postal system use, 90
scanning technology, 93, 94
science-industry interaction, 10
semiconductors, 93, 197
silicon limits, 201
structural unemployment impact, 109
surplus stores, 15
system trends, 53
technology, 91, 93, 114, 203
transmitters-receivers, 170
U.S.-Japan competition, 105, 107
x-ray lithography, 93, 188–189, 198
Electronics, natural sciences relationship, 3
Electronics revolution
communications, scientific and technologic, 56
development, 4, 12
dynamics, 3
history, 3, 8, 12
innovations, 3
medical diagnostic and monitoring equipment, 5
national defense development use, 81
optical communications, 7, 12
productivity increases, 8, 9
radio and television, 3, 5, 9, 10
research processes, 56
satellite technology, 5
science-society interaction, 8, 9
sensors, 4
opportunities for unskilled labor, 109
telecommunications, 8
transducers, 4
Elementary fiber link, optical components, 167, 168
Emissions control
computerized systems, 79
standards, 86
Energy conservation, semiconductor electronics, 29
Energy use, 3
Engineering developments, 13
Engineering schools, computer sciences, 28
Engineering studies, television instruction, 63–67
England, fibers for transmission studies, 167
Environmental control emissions monitoring, 79, 86
Environmental Protection Agency, 86
European Space Agency, 51

F

FCC (*see* Federal Communications Commission)
FET (*see* Field-effect transistor)
FORTRAN, 16, 60, 124, 136
Fabrication processes
modified chemical vapor deposition (MCVD), 167, 169
Fairchild, silicon diodes manufacture, 110
Federal Aviation Administration, 51
Federal Communications Commission, 97, 98, 100, 102
Federal communications regulations, 97
Federal Reserve Banks, computer batch systems, 31, 32
Fibers
cable cladding, 169
connectors and splicing, 169
short-lengths applications, 171
See also Glass fibers
Fiberguide system, 171

Field effect transistor
electron mobility, 201
hot electrons, 199
miniaturization limits, 199
Finite numbers, 189–190
Finite problem-solving, 196
Ford Foundation, 62

G

General Electric, beam-addressable MOS, 183
Glass fibers
characteristics, 167
digital transmissions capability, 168, 170
graded index, 168, 169
high-, low-loss fibers, 167, 169
light-wave loss, 167, 169
transmission lines use, 167
See also Fibers
Grid paths, 192, 194

H

Health care
CAT scanner, 41
electronic devices, 35
heart monitors, 36
See also Medical devices and Medicine
Hewlett-Packard, television instruction, 63–65
Human factors
engineering
electronics industry, 162
human-computer language problems, 162
human-machine interaction, 161, 165
information gathering, 161
job-equipment design, 162
World War II impact, 161
psychology
human performance studies, 164
industrial environment, 161
military environment, 161

I

IBM (*see* International Business Machines)
IMP-16, 143
IMPATT (*see* Impact avalanche and transit time)
INTELSAT (*see* International Telecommunications Satellite Organization)
Image reconstruction, 46
Impact avalanche and transit time, 52
diodes, 204
Implantable telemetry systems
cardiovascular research, 38
physiological measurements, 38
See also Medical devices
Industrial control systems, 76–77
costs, 78, 80
government requirements compliance, 77
hardware/software development, 78
history, 77–78
primary plant mission, 77
productivity improvement, 77
safe operating conditions, 77
technology development, 77–78, 80
Industrial plants, computerized process control, 76, 78, 80
Industrial Revolution, 116
electronics revolution comparison, 3
Information
management
firmware enhancements, 125
intelligent controllers, 125, 126
minicomputer back-end processors, 125, 127
modular data base machines, 125, 128
utility structure, 128
services
communication systems, 90, 91
U.S. gross national product, 7
U.S. work force, 7
systems
costs of humanizing, 154
economics, 154
hardware/software requirements, 122
humanizing criteria, 152
light-wave transfer, 7
management needs, 121
privacy, 154
procedures, 153
scarce resources, 121
service productivity, 120
social responses, 34
society complexity, 121
software, 122
Stanley House guidelines, 152–153
structure, 122
technology, 116
Injection lasers, materials requirements, 207
Instruction, television
computer-aided, 32
engineering studies, 63–67
Hewlett-Packard, television instruction, 63–65
international television, 55
technologies, 61, 62, 64, 122, 132
television instruction experiment, 63
videotaped instruction, 61, 64
See also PLATO and Tutored Videotaped Instruction
Insurance Corp. of British Columbia, auto-plan, 143–154
Integrated circuits, 4, 9, 12
cardiovascular research, 38
cascade charge coupled device, 37
chips, 19
dimensions, 19–20
information manipulation, 18
large-scale, 4, 12, 18, 81, 83, 93, 136, 144
markets, 26
microelectronics, 186
microwave, 52, 53
missile guidance systems, 82
neural prostheses, 39
planar technology, 197
production problems, 20
redundancy, 21
reliability, 204, 205
semiconductor memory chips, 39
sensory prostheses, 40
technology development, 85, 111
Integrated electronics
imaging devices in medicine, 37
implantable sensory devices, 40
prostheses, 38–39
Intel 8080, 145
point-of-sale market, 140
Intel 8748, 122
International Business Machines, 96, 97, 98, 102
IBM 360, 125
IBM 370, 125
IBM 650, 56, 57
IBM 704, 58, 82
IBM 705, 82
IBM 7094, 58
IBM OS/360, 156
computer industry, 6, 16, 17, 25, 26, 96, 97, 98, 102
digital satellite services, 102
semiconductor use, 25
International Telecommunications Satellite Organization, 47, 48–50, 52
active medium propagation (AMT) amplifiers, 52
antennas technology, 48
components, 52
earth stations communications, 48
global communications, 47, 49
high frequency operation, 52
international television, 55
lease transponder service, 50
organization and management, 50
participating countries, 49–50
technology development, 47

J

Japan
computer competition, 153
computer exports, 105
consumer electronics dominance, 108
data processing, 105–107, 158
economic policy, 105
manufacturing policies, 106, 107
Ministry of International Trade and Industry (MITI), 105, 106
Selfoc, 167
technology, 105, 107
television cost effectiveness, 107
television receivers exports, 107
U.S. industry competition, 105–106, 107, 108
Western Europe competition, 106
Josephson effect, 9, 185

L

Larams (*see* line-addressable RAMs)
LED (*see* Light-emitting diode)
LSI-11, 144
Large-scale integrated circuits, 4, 12, 18, 81, 83, 93, 136, 144
photolithography, 181
random access memory (RAM), 139
read only memory (ROM), 139
semiconductor memories, 179
Lasers
dark-line defects, 205
materials, 207
semiconductor injection carrier generator, 168
solid-state, 167
wave-length dispersion, 168
Latin squares, 191–192
Lawrence Berkeley Laboratory, 45, 46
Legislation, P.L. 94–163, 86
Light-emitting diode, 88
cost effectiveness, 167
dark-line defects, 205
multimode fibers, 167
optical carrier generators, 167
short-distance transmission, 167
single-mode fibers, 168
Light-wave transmission, 171
Lincoln Laboratories, MIT, 99
x-ray lithography, 188
Lincoln Memory Test Computer, 82
Line addressable RAMs, 182, 183
Lithography, miniaturization limitations, 197
Logic circuits, package design, 200
Lower atmosphere, kinetics models, 58

M

MARISAT, 51, 54, 55
MBE (*see* Molecular beam epitaxy)
MIC (*see* Microwave integrated circuit)
MIS (*see* Management Information Systems)
MISAR (Microprocessed sensing and automatic regulation), 86, 87
MOS (*see* Metal oxide semiconductors)

MTBF (*see* Mean time between failure)
MTD (*see* Moving Target Detection)
MTI (*see* Moving Target Indicator)
Mail (*see* Postal Service)
Management information systems, 17, 122, 151, 155
  Stanley House guidelines, 152–153
Marketing of goods, computers application, 33
Mathematicians, 191, 194–195
Mathematics calculations, 142, 144
Maxicomputers, 14
Mayo Clinic, 41, 42, 45
Mean time between failures, 181
Mechanical addressing, 185
Mechanical simulation (MECHSIM), 163, 164
Medical devices
  brain scanners, 41
  cardiac pacemaker, 39
  computerized tomographic scanner, 36, 41
  computerized x-ray scanners, 42, 44–46
  Doppler flowmeter, 38
  electronics, 5
  heart monitors, 36
  imaging systems, 36–37
  implantable telemetry systems, 38
  instruments, transducer elements, 37
  Optacon, 39
  sensory prostheses, 40
  transmodal sensory prostheses, 39
  ultrasonic blood flowmeter, 37–38
  ultrasonic imaging systems, 37
  visual prostheses, 40
  x-ray limitations, 41–42
Medicine
  electronic diagnostic devices, 35, 42, 44
  electronic invasive instruments, 36
  electronic monitoring instrumentation, 35
  health care, 35
Memories
  beam addressable MOS, 183
  capacity, 181
  electron beam-assessed, 183
  gap fillers, 182, 183
  optics, 184
  technology, 123
  thermal limitations, 201
  See also Core memories and Bubble memories
Mental processes theories, computations, 59
Metal oxide semiconductor, 180
  automotive microprocessor, 86
  FET units, 199
  high capacity memories, 180
  solid-state diffusion, 21
Meteorological measurements, 59
Michigan Educational Research Information Triad (MERIT), 133
Microcomputers, 122
  components, 139
  cost effectiveness, 97
  technology, 123
  U.S.-Japanese competition, 108
Microelectronics, 186
Microprocessor, 13, 14, 15, 22, 122
  communication satellite, 53
  data flow cycle, 147
  data management and processing, 149
  design, 139
  digital device, 146
  economics, 138, 139
  instrument control and data acquisition, 148, 149
  instrument operation, 146, 148, 149
  INTEL 8748, 122
  keyboard and display function, 147, 148
  LSI-11, 144
  Motorola M6800, 142, 145
  programming, 149, 150
  software, 144
  system development, 147–148
  technology, 123
Microwave integrated circuit, 52, 53
Minicomputers, 14, 15, 138
  central processors, 58
  microprocessors, 149, 150
  networks, 58
  visual displays, 58
Minuteman missile system, 25, 29
Missile guidance systems, 82
Modeling, atmospherics, 58, 59
Molecular beam epitaxy
  electron diffraction studies, 204
  laser devices, 203
  optical studies, 204
  process development, 204
  surface science research, 204
Motorola M6800, 142, 145
Moving Target Detection, 83, 84, 85
Moving Target Indicator, 83, 85

N

NASA (*see* National Aeronautics and Space Administration)
NCU (*see* Network control unit)
NEEMIS (*see* New England Management Information System)
NERComP (*see* New England Regional Computing Program)
NSF (*see* National Science Foundation)
National Aeronautics and Space Administration, 51, 53
National Cash Register, electronic registers, 33
National defense, electronic control systems, 83
National Science Foundation, 130
Network control unit, 131
Network Interface Message Processing Host (NIMPH), 136
New England Management Information System, 128
New England Regional Computing Program, 130
  educational programs, 132
  history and development, 129–130, 133
  microcomputer uses, 136
  multicomputer network, 130, 131
  on-campus/off-campus service problems, 133
  operations cost effectiveness, 132
  research programs, 132
  resource chaining, 134, 135
  services flow, 134
  supplier-distributor relationships, 132–133, 135
  switching techniques, 136
  tax status, 132
  terminal-oriented usage, 132
  transmission limitations, 132
New Jersey-New York atmospheric chemical model, 59
Noninvasive instruments, 37, 41
Northeast Academic Science Information Center (NASIC), 136
Numbers, finite and realistic, 189–190

O

OCR (*see* Optical character recognition)
OECD (*see* Organization for Economic Cooperation and Development)
On-line retrieval systems, BELLREL, 57
On-line service, banks, 31, 32
Operating Network Advisory Committee (ONAC), 134, 136
Optical
  carrier generators, 167, 168
  character recognition, 52
  communications, 7, 12
  memories, 184
Organization for Economic Cooperation and Development, 106
Organizational Government Advisory Committee (OGAC), 134, 135

P

PCM (*see* pulse-code modulation)
PETT (*see* Positron emission transaxial tomograph)
PL/1, 16, 124
PMP (*see* Parallel microprogrammed processor
POS (*see* Point-of-sale)
Packet switching, 100
Parallel microprogrammed processor, 84, 85
Personnel subsystem development (PSD), 165
Photolithography, 197
  computer circuits, 20–21
  large-scale integration, 181
  masks, production, 187
  silicon, 20–21
PLATO
  communication media, 73
  computer-based education, 68–70
  course components, 71
  course integration, 71
  instant course feedback, 71
  laboratory experiments simulation, 72
  lesson development costs, 74, 75
  lesson materials and procedures, 68, 72–73
  problem-solving, 69
  research use, 68
  science education use, 68
  self-paced instruction, 69
  system design, 74
  system operation, 75
  terminal displayed texts, 68–69, 73–74, 75
  TUTOR programming language, 72
Pneumatic transmitters, 78
Pneumoencephalography, 42
Point-of-sale terminals, 33, 34
Positron emission transaxial tomograph, 46
Postal Service
  business and government use, 9
  computer uses, 6
    ARPA computer network, 6
  correspondence, 90
  direct distance dial (DDD), 91
  electronic fund transfer, 30, 31, 90–91
  electronics technology use, 113
  history and development, 90
  magazines and newspapers, 90
  mail transactions, 90
  merchandise orders, 90
  remote transmissions, 92, 96
  telephone network, 91
  volume, 92
  See also Electronic mail system
Printing technology, 93, 94
Privacy, computer industry, 119
Processing technology, 97, 116
Processors, 143
Programming
  computer software, 16
  "Go to" procedures (GOTO), 159
  languages, 16
  mathematics, 192
  structured, 158–159
  techniques, 158, 159, 160
Project Relay, 47, 48
Project Telstar, 47, 48
Prostheses, 38–40
Psychology, computerized research, 59, 120
Publishers, computer economics, 57

Pulse-code modulation, 170
Pulse-position modulation (PPM), 170–171

Q

Quantile-quantile display, 59

R

Radar
  coherent processing interval, 84
  microwave frequencies, 81
  pulse repetition frequency, 84
  target detection devices, 84
Radiation Laboratory, MIT, 81
Radiography, low-dosage imaging, 46
Radionuclide imaging, 46
Radionuclide scanning, 42
Random access memory (RAM), 147, 148, 149, 150, 179, 182, 183
  access time comparisons, 182
  fabrication, 186
Read only memory (ROM), 147, 148, 149, 150
Reservations services, 33
Retail trade computers, 33, 34

S

SAGE (*see* Semi-Automatic Ground Environment)
SAW (*see* surface acoustic-wave device)
Santa Rosa, Calif., television instruction experiment, 63
Satellite communications
  antenna design, 52
  Canada, 50
  domestic systems, 50
  Indonesia, 50
  mobile systems, 51
  newspapers, transmission, 55
  space technology, 51
  technology, 5, 6, 10
  tracking radar, 82
  transponder components, 51
  tripartite organization, 51
  United States, 50
  See also COMSAT
Satellites
  domestic market, 102
  INTELSAT, 1, 10
  NAVSTAR, 5
  Telstar, 10
Scanners
  brain, 41
  charge-coupled devices (CCD), 93
  computer improvement, 46
  computerized tomographic, 36, 41
  computerized x-ray, 42, 44, 46
  electromagnetic interference, 43
  optical scanning terminal, 92
  radionuclide, 42
  technology, 93, 94
Scanning electron beam system, 187, 188
Semi-Automatic Ground Environment, 82, 163
Semiconductor
  electronics, 23
    energy conservation, 29
    history and development, 23, 24
    markets, 25
    metal oxide semiconductor, 21, 26, 198, 199
  military uses, 23
    research and development, 27–24
    space uses, 23
    telecommunications industry, 29
  industry
    business ventures, 26–27
    circuitry, 26
    economics, 23–24
    government funding, 23–25
    memory chips, 26
    microprocessors, 26
    Minuteman missile system, 25, 29
    sales, 111
Semiconductors
  charge packets, 172
  chip density limits, 200
  chip packaging, 200
  chip size and integrity, 197
  defect motion, 205
  digital integrated circuits development, 201
  dopants, 168, 187
  history and development, 179–180
  industry, 23, 25
  ion implantation, 187
  manufacture techniques, 187
  memories, large-scale integration, 179
  microelectronics, 197
  PIN, 167, 168, 171
  polycrystalline thin films, 203
  RAM capacity, 181
  silicon technology, 201
  storage, 183
  technology, 23, 27, 202
  voltage levels, 198
Sensory prostheses, 40
Sequential processing
  abstract machines, 160
  axiomatic basis, 159
  function algebra, 159
  predicate transformers, 159
  software design, 158–159, 160
  structured programs, 159, 160
  synchronization, 158, 160
Signal processors
  characteristics, 173
  charge-transfer device, 172
  surface acoustic-wave device, 172
  technology, 172–173
Silicon chips, 180, 181
Software
  abstract machine systems, 158
  complexity problems, 156
  computer programs, 16
  costs, 16
  definitions, 156
  design hierarchy, 160
  development, 155, 156
  engineering, 16, 156, 158
  human-machine operation, 156–157, 165
  IBM OS/360, 156
  quality control, 16
  mathematics, 155
  sequential processes design, 158, 160
  system organization, 160
Software-hardware interface, 157–158
Solar power, semiconductor electronics, 29
Solid-state electronics, 109
  defects, 205
  diffusion, 21
  laser research, 207
  materials research, 206–207
  molecular beam epitaxy (MBE), potential use, 204
  optics research, 207
  optoelectronics reliability, 205
  scanners, charge-coupled devices (CCD), 93
  superconducting materials, 206
  surface science, 203, 204
  switching speeds, 206
  technology development, 202, 203, 206–207
Solid-state physics, 167
  antenna systems, signal processing, 48
  electronics developments, 4, 9
  technology advancements, 54
Soviet Union (*see* U.S.S.R.)
Space Shuttle, 53
Space stations communications, geostationary orbit, 47
Spacecraft, attitude control, 52
Stanley House information system guidelines, 152–153
Stein tree
  algorithms, 194
  electrical connectors, 194, 196
Storage-capacity costs, 184–185
Storage tubes, 183
Students
  engineering television instruction, 65–66
  PLATO lessons, 68–69
Surface acoustic-wave device
  band pass filters, 177
  elastic wave energy propagation, 172
  electrode transducers, 174
  frequency detection, 176
  principles, 175
  processing signals, 177
  research, 172
  signal processing costs, 178
System
  computerization, 151, 155
  design
    computer-based directory assistance system, 163, 164
    hardware-software development, 162, 165
    human error, 163, 164
    human-machine interface, 163
    human performance problems, 163, 164
    instructional technology, 163
    personnel subsystems, 162, 165

T

TELESAT, 50, 54
TVI (*see* Tutored Videotaped Instruction)
TWT (*see* Traveling wave tube amplifiers)
Technology transfer, 107
Telecommunication
  antitrust suits, 101
  coaxial cables, 171
  fiber and cable fabrication, 168–169
  glass fiber use, 167
  intercity, 167, 171
  investments, 101
  management, 122
  markets, 100, 102
  modulation format, 170
  monopolies, 103
  multiple interindustry competition, 102
  regulations, 100
  research and development, 101, 102
  switching systems, 112
  system optics, 167
  technology innovations, 102
  trends, 104
  wire pairs, 171
Telephone industry
  American Telephone and Telegraph Co., 101
  computer-based directory assistance, 163, 164
  equipment, 101, 103
  organization, 101, 103
  public policy debates, 102–103, 104
  service, 100, 101, 103
Teleprinter systems, 90
Television
  industry, competition, 101, 103
  manufacture, U.S.-Japan competition, 105, 107

Time-sharing, UNIX system, 60
Tracking radar, 83, 84
Trade Act of 1974, 106
Transducers, 27
Transistor electronics, military uses, 23
Transmodal sensory prosthesis, Optacon, 39
Transversal filtering, 176
Traveling wave tube amplifiers, 48, 51–52
TUTOR programming language, 72
Tutored Videotaped Instruction
  experiments, 63–65
  format, 67
  off-campus studies, 62, 64
  on-campus/off-campus student averages, 65
  Santa Rosa, Calif., experiment, 63

U

UNIX time-share system, 60
Ultrasonic imaging system, 36–37
Ultrasound application, imaging, 37
Universal product coding (UPC), 34
Universities, engineering research, 28
Urban atmosphere computations, 58
Users Services Advisory Committee (USAC), 134, 135
U.S. International Trade Commission, 106
U.S. manufacturing, computer industry, 107
U.S.S.R.
  atomic device development, 81
  computer competition, 157–158

V

Vacuum tube components, industry, 26
Videotaped instruction, 61, 64
Visual prostheses, technology problems, 40
von Neumann, John, 12, 13, 15, 17

W

Western Electric Co., 111, 112
WHIRLWIND, 82
Whole body scans, 44
  See also Medical devices and Medicine

X, Y, and Z

X-ray lithography
  diffraction advantages, 188–189
  resolution limits, 198
  techniques, 188
X-ray scanners, 42, 43, 44–45, 46
X-ray source geometry, fan beam, 44
X-ray tomography, computer-aided device, 36
Zenith Radio Corporation-Japan antitrust case, 106

BECOME A PART OF THE

American Association for the Advancement of Science

Founded in 1848, AAAS is the world's leading general scientific society with 113,000 individual members interested in the advancement of science, in improving the effectiveness of science in promoting human welfare, and in increasing the public understanding and appreciation of science. Through its nearly 300 affiliated societies and academies covering the entire spectrum of science and technology and representing nearly 6 million individuals directly involved in science and technology, AAAS has a broad base of expertise for its continuing programs and special projects. Through its membership and its affiliates, AAAS exercises leadership in the analysis of the technological, social, and political ingredients in significant problems facing society today.

COME JOIN US NOW

For information about becoming a member of the Association, write to

AAAS Membership Department
1515 Massachusetts Ave., N.W.
Washington, D.C. 20005

WHAT AAAS CAN DO FOR YOU

AAAS keeps you in tune with the latest scientific developments through:

- SCIENCE, the weekly magazine offering definitive articles and up-to-the-minute reports on topics and issues about which you **must** know.
- the Annual National Meeting with symposia and lectures on recent developments in science, and informed discussions on policy issues about which you **should** know.
- the quarterly review magazine SCIENCE BOOKS & FILMS, the Science Book Lists, and the Science Film Catalog to help you select the best science books and films because you **want** to know.
- important publications like the Science Compendia on energy, food, population, and materials, the many audiotape cassettes and albums, and the published symposia which result from the Annual Meeting, which offer you a broader perspective because you **need** to know.

AAAS gives you the means to influence important decisions and processes by:

- giving national and regional policy-makers the science facts they need through special seminars and the Congressional Fellows Program.
- providing forums on such problems as scientific freedom and responsibility, the legal and technical implications of whether modification, the implications of energy development in the west, and more.
- relaying reliable science information to the news media.
- promoting public understanding of science and improving science curricula in the schools.
- improving international cooperation among scientists through innovative ventures like the new inter-American trilingual journal INTERCIENCIA.
- expanding the opportunities available to minorities, women, and the handicapped in all fields of science.
- joining with 113,000 others, all of whom have a vital interest in science and society.

The potential for developing inexpensive processing power is, truly, awesome.

ROBERT N. NOYCE "Large-Scale Integration—What Is Yet to Come?"

It is now technologically feasible to replace defective nervous tissue with synthetic neural circuits of similar size and power needs.

R. L. WHITE AND M. D. MEINDL "The Impact of Integrated Electronics in Medicine"

We are now in the early stages of a revolution in processing information that shows every sign of being as fundamental as the earlier energy revolution.

HERBERT A. SIMON "What Computers Mean for Man and Society"

From a fundamental point of view, our present degree of control over electrons and their motions in solids may be compared to our ability to manipulate light at the beginning of the 19th century.

J. A. GIORDMAINE "Solid-State Electronics: Scientific Basis for Future Advances"

My belief based on the experience of us all is that the balance of power and the ratio of intelligence between man and computer is still indeterminate.

RUTH M. DAVIS "Evolution of Computers and Computing"